N° d'ordre NNT : 2021LYSE2025

THESE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de

L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 486

Sciences économiques et de gestion

Discipline : Sciences économiques

Soutenue publiquement le 2 juillet 2021, par :

Claire RIMBAUD

Three Essays on Guilt Aversion

Theory and Experiments.

Devant le jury composé de :

Loukas BALAFOUTAS, Professeur d'université, Université d'Innsbruck, Président

Astrid HOPFENSITZ, Maître de conferences HDR, Toulouse School of Economics, Rapporteure

Martin DUFWENBERG, Professeur d'université, University of Arizona, Rapporteur

Brice CORGNET, Professeur, EM Lyon Business school, Examinateur

Marie-Claire VILLEVAL, Directrice de recherche, CNRS, Directrice de thèse

Contrat de diffusion

Ce document est diffusé sous le contrat *Creative Commons* « **Paternité – pas d'utilisation commerciale - pas de modification** » : vous êtes libre de le reproduire, de le distribuer et de le communiquer au public à condition d'en mentionner le nom de l'auteur et de ne pas le modifier, le transformer, l'adapter ni l'utiliser à des fins commerciales.

THÈSE DE DOCTORAT EN SCIENCES ECONOMIQUES

Présentée en vue de l'obtention du grade de docteur de l'Université de Lyon,

délivré par l'Université Lumière Lyon 2

THREE ESSAYS ON GUILT AVERSION: THEORY AND EXPERIMENTS

Soutenue par

CLAIRE RIMBAUD

Dirigée par :

MARIE CLAIRE VILLEVAL - Directrice de recherche CNRS, GATE (FR)

Soutenue le 2 juillet 2021

Devant le jury composé de :

LOUKAS BALAFOUTAS	-	Professeur, University of Innsbruck (AU)	Examinateur
BRICE CORGNET	-	Professeur, EM Lyon Business School (FR)	Examinateur
MARTIN DUFWENBERG	-	Professeur, University of Arizona (USA)	Rapporteur
ASTRID HOPFENSITZ	-	Maitre de Conférences (HDR), Toulouse School of Economics (FR)	Rapportrice

---

Université de Lyon - Ecole Doctorale Sciences Ecomiques et Gestion

Université Lumière Lyon 2

Groupe d'Analyse et de Théorie Economique, Lyon Saint Etienne

University Lumière Lyon 2 does not give any approbation or disapprobation about the thoughts expressed in this dissertation. They are only the author's ones and need to be considered as such.

Remerciements

Je tiens à remercier Marie Claire Villeval. J'ai choisi de venir à Lyon pour pouvoir faire mon mémoire avec vous. C'est grâce à votre soutien et vos conseils que je suis partie faire mon master à Nottingham. Et c'est pour les retrouver que je suis revenue faire ma thèse à Lyon sous votre direction. Je vous suis extrêmement reconnaissante pour votre humanité, votre écoute mais aussi votre pragmatisme et votre expérience de chercheuse. Vous avez su adapter le ton de votre encadrement aux différents moments qui constituent une thèse, entre l'excitation et le découragement.

Je remercie mon co-auteur, Giuseppe Attanasi, pour ses conseils et ses encouragements autant que pour sa bonne humeur. C'est grâce à toi que j'ai pu entrer dans le monde de la théorie des jeux psychologiques et je t'en suis reconnaissante.

Je remercie Loukas Balafoutas, Brice Corgnet, Martin Dufwenberg et Astrid Hopfensitz d'avoir accepté de participer à mon jury de thèse.

Je remercie Lata Gangadharan de m'avoir encadrée à Monash Université pendant près de 7 mois. Pour les discussions très enrichissantes que j'ai pu avoir au sein du département d'économie, je remercie également Zhengyang Bao, Behnud Mir Djawadi, Nick

Feltovich, Ben Grodeck, Philip Grossman, Tomas Zelinsky et Yves Zenou. Malgré la distance avec la France, j'ai passé un excellent séjour grâce aux personnes qui m'y ont accueillie avec beaucoup de gentillesse au sein de l'université et à l'extérieur.

Je remercie tous les membres du GATE, les chercheurs et le personnel administratif. Quentin mérite une mention toute particulière car sans lui, aucune expérience ne pourrait avoir lieu. Je suis fière de compter parmi mes amis Alice, Charlotte, Clément, Jocelyn, Julien, Liza, Marius, Maria, Maxime, Morgan, Rémi, Sorravich, Valentin et Vincent.

Je suis extrêmement chanceuse d'avoir eu des soutiens tout au long de ma thèse, mais aussi tout au long de mon cursus scolaire. Je remercie Claire et Romain de me rappeler le chemin parcouru. Je remercie Bastien, Pierre et Thomas pour leur fidélité, leur bonne humeur. Je remercie Abigail, Julie, Justine et Pauline d'être toujours là et de former ce cercle d'amitié si sécurisant. Je remercie évidemment les lyonnais Jérémy, Marius, Sarra, Steve, Tais et Tatiana parce que la vie sans eux n'aurait pas été la même. Je remercie mes colocs Cécile et Lisa qui ont suivi mes déboires de thésarde au jour le jour. Tous à votre manière vous avez su m'inspirer, me soutenir, me rabrouer ou me faire rire et me prendre dans vos bras, et s'il y avait un diplôme d'amitié vous l'obtiendrez avec les félicitations.

Je remercie Thomas d'avoir su me soutenir dans les moments difficiles et m'accompagner dans les moments plus festifs. Ton amour au quotidien m'est précieux.

Enfin, je remercie ma famille. Je pense à mes grand-pères qui seront, j'espère, fiers de moi et à mes grands-mères qui m'ont accompagnée de leur bienveillance. Je suis infiniment reconnaissante à mes parents pour leur soutien indéfectible.

Résumé de la thèse

Motivation de la thèse

Préférences dépendantes des croyances

Bentham a été le premier à introduire le concept de fonction d'utilité en 1789, un concept qui est aujourd'hui au cœur de la science économique. Il considérait que les individus cherchent à maximiser les plaisirs et à minimiser les souffrances. Il est intéressant de noter que les plaisirs et les souffrances peuvent correspondre à des croyances (par exemple, le plaisir et la souffrance liés à la mémoire, ou le plaisir et la souffrance liés à l'imagination). Cependant, depuis cette première contribution, les économistes ont considéré que les individus cherchent à maximiser leur gain attendu et ont surtout envisagé les croyances comme une contrainte sur ce gain. Ce n'est que récemment que les économistes expérimentaux sont revenus à une conception plus large de ce que les individus cherchent à maximiser, en remettant en question l'hypothèse d'un agent exclusivement intéressé à son propre gain et en intégrant dans la fonction d'utilité des préférences qui tiennent compte d'autres facteurs (pour une synthèse, voir Cooper, 2009). Pourtant, la plupart de ces nouveaux modèles (par exemple, Fehr and Schmidt, 1999; Charness and Rabin, 2002) continuent de définir les préférences basées sur les conséquences monétaires des choix (par exemple, aversion à l'inégalité, souci d'efficacité). En allant dans le sens de la suggestion de Bentham, il est crucial de considérer également les préférences basées sur

les croyances. Par exemple, l'aversion à la culpabilité, objet d'étude de cette thèse, est un type de préférence qui conduit les individus à éviter de décevoir leurs croyances sur les attentes des autres. Cette préférence basée sur les croyances affecte une variété de décisions. Ainsi, l'essor des politiques respectueuses de l'environnement peut en partie s'expliquer par la culpabilité des décideurs politiques face à l'expression des attentes des jeunes générations. En outre, l'influence de l'aversion à la culpabilité contribue à éclairer le fossé entre certains pays en développement dans lesquels la corruption est courante et d'autres où demander un pot-de-vin est une exception. Dans le premier cas, demander un pot-de-vin est dénué de toute culpabilité puisque cela ne déçoit pas les attentes en matière d'intégrité, et donc la corruption persiste. Dans le second cas, l'aversion à la culpabilité empêche la corruption car les attentes en matière d'intégrité sont élevées. Dans notre vie quotidienne, décider du montant du pourboire à donner à un chauffeur de taxi dépend également de la culpabilité que nous pouvons ressentir en décevant les attentes du chauffeur de taxi.

Au delà de la seule aversion à la culpabilité, l'ensemble des préférences basées sur les croyances peut aussi aider à comprendre pourquoi les décideurs politiques ne doivent pas se concentrer exclusivement sur la richesse matérielle d'une population pour évaluer son bonheur (Easterlin, 1995). En effet, le bien-être des individus peut dépendre de la manière dont ils se perçoivent (et dont les autres les perçoivent), de l'évaluation de leurs choix par rapport à leurs attentes initiales, de la manière dont ils valorisent la réciprocité des intentions des autres ou de la manière dont ils ressentent leurs émotions. Chacun de ces motifs peut être modélisé en incorporant les croyances en tant que préférences dans la fonction d'utilité.[1]. Les études récentes de Battigalli and Dufwenberg (2020) et

[1]Les outils permettant d'explorer les préférences dépendantes des croyances ont été développés dans le cadre de la théorie des jeux psychologiques. Ce cadre a été initialement proposé par Geanakoplos et al. (1989) et étendu par Battigalli and Dufwenberg (2009). La principale caractéristique de ce cadre théorique est de laisser l'utilité à un noeud final de décision dépendre des croyances, alors que ce n'est pas le cas

Loewenstein and Molnar (2018) illustrent à quel point les préférences dépendantes des croyances sont répandues. Premièrement, ces préférences permettent de rendre compte de l'importance des préoccupations liées à l'image que les individus tentent de maintenir. Cette image est liée soit à leurs actions (je veux que les autres croient que j'ai fait une action X), soit à leurs traits de caractère (je veux que moi-même/les autres croient que je suis Y). Par exemple, un individu peut ne pas aimer que les autres pensent qu'il a triché (Dufwenberg and Dufwenberg, 2018) ; il peut aussi ne pas aimer être perçu comme intéressé (e.g., Bénabou and Tirole, 2006 ; Grossman and Van Der Weele, 2017). De manière générale, cette littérature remet en cause l'idée que les actions pro-sociales soient motivées par de pures préférences pro-sociales, mais montre qu'elles répondent également aux croyances du décideur liées à l'image donnée. Une autre illustration est donnée par le modèle de Mannahan (2019), cité par Battigalli and Dufwenberg (2020), qui suppose que l'utilité d'un individu augmente lorsqu'il croit que ses capacités (par exemple, son intelligence) sont élevées. Ses préférences pour de telles croyances peuvent conduire un individu qui va subir une évaluation extérieure à se saboter avant cette évaluation afin de rendre le signal sur ses capacités sans valeur (par peur de découvrir ses véritables capacités).

Deuxièmement, les fonctions d'utilité dépendantes des croyances autorisent des préférences dépendantes des références basées sur les attentes, c'est-à-dire des préférences qui dépendent de l'attente initiale du décideur (pour une synthèse, voir O'Donoghue and Sprenger, 2018). L'un des exemples les plus marquants de telles préférences est donné par les modèles de Kőszegi and Rabin (2006) et Kőszegi and Rabin (2007).[2] Ils ont proposé que les individus aient une aversion à la "déception", c'est-à-dire pour l'obtention d'un

dans la théorie des jeux traditionnelle. On observe une augmentation constante du nombre d'articles citant ce cadre théorique entre 1991 et 2017 (Azar, 2019).

[2]Des travaux séminaux dans la même veine ont été menés par Loomes and Sugden (1982) et Bell (1985).

gain plus faible que prévu. Ce cadre permet de rendre compte de divers phénomènes tels que le niveau de revenu quotidien visé (Farber, 2005), les choix des consommateurs (Heidhues and Kőszegi, 2008), l'effet de dotation (Knetsch and Wong, 2009), l'effort réel dans la compétition (Gill and Prowse, 2012) ou la recherche d'emploi (DellaVigna et al., 2017).

Troisièmement, le cadre dépendant des croyances permet aux chercheurs de modéliser les préférences réciproques basées sur l'intention (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004, 2019). Ces modèles soutiennent que les individus aiment répondre à la gentillesse/méchanceté par la gentillesse/méchanceté, et que la valence d'une action est évaluée par rapport aux intentions. Les travaux appliqués sur les préférences réciproques montrent que ces préférences peuvent expliquer des phénomènes aussi divers que les effets de formulation (Dufwenberg et al., 2011), les différends commerciaux (Conconi et al., 2017), la communication (Le Quement et al., 2018), ou les négociations sur le changement climatique (Nyborg, 2018).

Enfin, en prenant en considération les préférences dépendantes des croyances, les chercheurs ouvrent la voie à l'introduction des émotions dans les fonctions d'utilité. Comme l'a noté Elster (1998), les économistes ont largement négligé l'étude des émotions. Il soutient que "*les émotions sont déclenchées par les croyances*" (p.49), ce qui fait des fonctions d'utilité basées sur les croyances leur terrain d'action naturel. L'impact des émotions est double. D'une part, les émotions peuvent conduire à une réponse comportementale (c'est-à-dire à une action). Par exemple, sur la base de l'hypothèse de frustration-agression de Dollard et al. (1939), Battigalli et al. (2019) ont proposé un modèle de colère et de frustration. Ils considèrent que les individus qui sont frustrés par leur gain (par rapport à leurs attentes) réagissent avec colère et peuvent être prêts à s'engager dans une action de vengeance (destruction du gain des autres). D'autre part, les

émotions peuvent également influer sur le comportement dans la mesure où les individus anticipent les émotions (positives ou négatives) que vont déclencher leur comportement et en tiennent compte lors de la planification. Par exemple, l'anxiété que le médecin anticipe chez son patient peut affecter la décision du premier de révéler ou non, un diagnostic médical défavorable (Caplin and Leahy, 2004). Comme cité précédemment, l'anticipation de la déception, émotion négative, peut influencer le comportement présent (Loomes and Sugden, 1982; Bell, 1985). Nous considérons maintenant de manière plus approfondie l'émotion au centre de la présente thèse, à savoir la culpabilité.

Aversion à la culpabilité

En psychologie, Baumeister et al. (1994) a distingué deux fonctions de la culpabilité. Premièrement, la culpabilité étant aversive, son *anticipation* va décourager les comportements nuisibles. Deuxièmement, lorsque le mal est fait, l'*expérience* de la culpabilité encourage le transgresseur (celui qui a deçu les attentes d'autrui) à restaurer une bonne relation avec l'autre partie et à éviter de répéter le même comportement nuisible à l'avenir.

En économie, Battigalli and Dufwenberg (2007) ont introduit le concept d'aversion à la culpabilité qui correspond à une anticipation aversive de la culpabilité.[3] Un décideur a

[3]Les psychologues ont largement concentré leurs recherches sur l'expérience de la culpabilité. Pourtant, des exceptions notables existent pour montrer que l'anticipation de la culpabilité peut modifier les comportements futurs. Par exemple, au moment de prendre la décision de donner ou non de la moelle osseuse, le fait d'envisager de futurs sentiments de culpabilité liés à l'inaction peut conduire à faire un don ; la culpabilité anticipée est le plus fort prédicteur de l'intention de faire un don d'organes après contrôle des normes, de l'auto-efficacité et de la discussion familiale. Dans un autre domaine, Wang and McClung (2012) a montré que la culpabilité anticipée influence significativement les intentions de téléchargement illégal des collégiens parmi ceux qui l'ont déjà fait au cours des 6 mois précédents (mais pas parmi ceux qui ne l'ont pas fait). En outre, les adolescents qui ont déclaré une culpabilité anticipée élevée à l'égard du comportement agressif étaient, selon leurs pairs, plus susceptibles de se comporter de manière prosociale et moins susceptibles de se comporter de manière antisociale (Olthof, 2012). Enfin, Steenhaut and Van Kenhove (2006) ont constaté que le renforcement de l'anticipation de la culpabilité encourageait les intentions des consommateurs à acheter des produits éthiques (sur la prise de décision éthique, voir également Motro et al., 2018).

de l'aversion à la culpabilité envers un autre joueur s'il subit une désutilité en décevant les attentes de l'autre joueur. Ceci est conforme à la proposition de Baumeister et al. (1994) qui stipule que "*Si les gens se sentent coupables de blesser leur partenaire [...] et de ne pas répondre à leurs attentes, ils modifieront leur comportement (pour éviter la culpabilité) d'une manière qui semble susceptible de maintenir et de renforcer la relation.*" Il est important de noter que le modèle de Battigalli and Dufwenberg (2007) prend en compte les attentes empiriques que porte autrui sur la manière dont le décideur *va* se comporter et non les attentes normatives sur la manière dont le décideur *doit* se comporter.[4]

La culpabilité se manifeste à de nombreux moments de notre vie. Les résultats de Baumeister et al. (1995) révèlent, qu'au cours d'une semaine de leur vie quotidienne, les participants ressentent de la culpabilité durant 13% de leur temps d'éveil. En économie, l'existence de l'aversion à la culpabilité a été évaluée dans une variété de jeux. Il a été démontré que l'aversion à la culpabilité favorise la coopération dans les jeux de confiance (trust game ; par exemple, Charness and Dufwenberg, 2006 ; Reuben et al., 2009 ; Bellemare et al. (2017) ; voir Cartwright, 2019b pour une synthèse) et dans les jeux d'envoi-réception (sender-receiver game ; par exemple, Battigalli et al., 2013). Des preuves d'un comportement d'aversion à la culpabilité ont été trouvées notamment dans un jeu de "biens de confiance" (credence goods game) (Beck et al., 2013), dans un jeu du "porte-monnaie perdu" (lost wallet game) (Dufwenberg and Gneezy, 2000) et dans un jeu de bien public (Patel and Smith, 2019). L'aversion à la culpabilité motive

[4]Cette distinction est liée à la différence entre le modèle de d'aversion à la culpabilité et les modèles de normes sociales tels que d'Adda et al. (2016). Cependant, Hauge (2016) et Krupka et al. (2017) ont montré que les attentes injonctives et empiriques comptent pour expliquer le comportement dans, respectivement, un jeu de dictateur et des accords informels. Dans le domaine des attentes empiriques, Danilov et al. (2019) a distingué l'action communément choisie par les autres décideurs dans les mêmes situations (norme sociale) et l'action attendue par une autre partie directement affectée par le choix du décideur (aversion à la culpabilité). Lorsque les informations sur les deux normes étaient révélées simultanément, les auteurs ont constaté que les deux types d'informations affectaient les transferts dans le jeu du dictateur

également le comportement prosocial dans les jeux de dictateur (dictator game ; par exemple, Ockenfels and Werner, 2014a ; Hauge, 2016 ; Bellemare et al., 2018) : plus le destinataire s'attend à recevoir, plus le dictateur donne.

De plus, l'aversion à la culpabilité a été proposée pour rendre compte de décisions dans des contextes plus divers que ceux correspondant aux jeux économiques habituels présentés ci-dessus. Dans un travail pionnier, Dufwenberg (2002) a étudié la situation suivante : une femme décide d'abord de soutenir ou non son mari (par exemple pendant qu'il étudie), puis le mari (qui a maintenant un diplôme) décide de divorcer (en récoltant tous les gains) ou de rester marié (en partageant les gains). Dufwenberg (2002) a montré que, si la sensibilité à la culpabilité du mari est suffisamment forte, le signal de confiance de la femme ayant soutenu son mari induit suffisamment de culpabilité pour "forcer" le mari à rester marié. Balafoutas (2011) a étudié un jeu entre un bureaucrate qui peut être corrompu, un lobby qui cherche à corrompre le bureaucrate et le public qui souffre d'une corruption éventuelle, où le bureaucrate peut se sentir coupable de décevoir les attentes du public. Dans une interaction ponctuelle, autoriser le bureaucrate à être averse à la culpabilité réduit la corruption. Cependant, lors d'interactions répétées, et sous certaines conditions (par exemple, lorsque les croyances sont mises à jour relativement rapidement), la société peut être piégée dans des croyances de corruption auto-alimentées. En outre, en examinant les conséquences de l'aversion à la culpabilité dans un contexte d'évasion fiscale, Dufwenberg and Nordblom (2018) ont supposé qu'un contribuable peut souffrir de culpabilité suite à une évasion fiscale. Ils ont montré que, lorsqu'on endogénéise le comportement de l'autorité fiscale, l'introduction de l'aversion à la culpabilité chez les contribuables réduit le taux d'inspection (ce qui implique une économie de fonds publics). Enfin, Cartwright (2019a) a souligné l'importance de l'aversion à la culpabilité dans la décision d'utiliser des drogues dans le sport. Il a constaté qu'avec une sensibilité à la culpabilité suffisamment élevée, les athlètes sont incités à courir sans

substance dopante si les autres s'attendent à ce qu'ils le fassent, indépendamment des comportements des autres athlètes de la course. Il convient de noter que ces travaux sont des modèles théoriques appliqués qui doivent encore être testés empiriquement.

Deux autres phénomènes bien établis, le respect des promesses et le favoritisme de groupe, ont été traditionnellement expliqués par l'aversion à la culpabilité. L'explication du respect des promesses basée sur les attentes propose que le fait de faire une promesse augmente les attentes du bénéficiaire, et que le décideur ne veut pas décevoir les attentes du bénéficiaire (Charness and Dufwenberg, 2006; Ederer and Stremitzer, 2017). De même, le favoritisme envers les membres de son groupe pourrait être expliqué, du moins en partie, par l'aversion à la culpabilité (plutôt que par les seules préférences intrinsèques du groupe). La culpabilité induite par les attentes des membres du groupe diffère de celle des membres de l'extérieur, soit parce que ces attentes sont considérées comme plus élevées par le décideur (Güth et al., 2009; Guala et al., 2013; Ockenfels and Werner, 2014b), soit parce qu'elles sont plus valorisées par le décideur (Morell, 2019).[5]

La culpabilité est prévalente dans de nombreux types de situations et a été proposée comme explication d'une variété de phénomènes. Pourtant, la question des facteurs situationnels qui modulent son influence reste peu étudiée (Ghidoni and Ploner, 2020).[6] À la suite de Balafoutas and Fornwagner (2017), la question n'est plus de savoir si l'aversion à la culpabilité existe, mais dans quelles circonstances elle est importante.

[5]Cependant, des études expérimentales ont suggéré que l'aversion ne parvient pas à rendre compte du respect des promesses (Vanberg, 2008; Di Bartolomeo et al., 2018) ou bien du favoritisme envers les membres de son groupe (Ciccarone et al., 2020) .

[6]Deux exceptions sont notables. Premièrement, il existe des preuves convergentes que les attentes " raisonnables " sont plus susceptibles d'être prises en compte par les joueurs averses à la culpabilité (Regner and Harth, 2014; Khalmetski, 2016; Balafoutas and Fornwagner, 2017; Danilov et al., 2019). Deuxièmement, Khalmetski (2016) et Bellemare et al. (2018) ont montré que la sensibilité moyenne à la culpabilité des dictateurs diminuait en fonction du niveau des enjeux monétaires.

Objectifs de la thèse

L'objectif de cette thèse est de mieux cerner la sphère d'influence de l'aversion à la culpabilité en étudiant (i) la direction de la culpabilité : l'aversion à la culpabilité influence-t-elle les individus, même lorsque leurs actions n'affectent pas les gains de la personne envers laquelle ils peuvent se sentir coupables, ou bien seulement lorsque leurs actions ont des conséquences monétaires directes pour cette personne ? (ii) certaines conditions nécessaires à l'émergence de l'aversion à la culpabilité : est-ce que la vulnérabilité de la personne envers laquelle les individus peuvent se sentir coupables influence l'émergence de la culpabilité ? (iii) la robustesse de l'aversion à la culpabilité face aux biais égoistes : dans quelle mesure les individus sont-ils stratégiques dans leur acquisition d'informations sur les attentes d'autrui (afin d'éviter de ressentir de la culpabilité) ?

D'une part, afin d'évaluer si la sphère d'influence de la culpabilité est plus large qu'on ne le pensait initialement, nous avons du remettre en question l'hypothèse du modèle de Battigalli and Dufwenberg (2007) qui soutient que l'on ne se sent coupable qu'envers les co-joueurs qui sont affectés monétairement. Sur le plan expérimental, nous avons introduit différents jeux à trois joueurs jusqu'alors jamais étudiés dans la littérature (Chapitres 1 et 2). D'autre part, afin de tester la nature "objective" ou "illusoire" de l'aversion à la culpabilité, nous avons placé les participants dans une situation où ils ne connaissaient pas les attentes de leurs co-joueurs et nous avons formalisé la manière dont les individus averses à la culpabilité devraient acquérir cette information (Chapitre 3).

Au chapitre 1, nous avons examiné une situation dans laquelle un donateur pouvait envoyer de l'argent à un bénéficiaire. Comme dans le cas des dons de charité ou des transferts gouvernementaux vers les pays en développement, cet argent devait être transmis par un intermédiaire qui pouvait en détourner une partie. Le comportement de cet

intermédiaire était au centre de nos recherches. Alors que les études précédentes ont exclusivement considéré la culpabilité potentielle de l'intermédiaire envers le bénéficiaire (comme dans les jeux de dictateur), nous avons pensé qu'un intermédiaire pouvait également éprouver de la culpabilité envers le donateur qui a renoncé à une partie de sa dotation pour augmenter le gain du bénéficiaire. Notre objectif de recherche était double. Premièrement, nous voulions documenter l'existence d'une aversion à la culpabilité chez les intermédiaires, tant envers le bénéficiaire qu'envers le donateur. Deuxièmement, nous voulions vérifier si la direction de la culpabilité, envers le donateur ou envers le bénéficiaire, affectait la prévalence ou l'intensité de l'aversion à la culpabilité. Ces objectifs ont exigé que nous proposion une extension du modèle de Battigalli and Dufwenberg (2007) pour permettre à l'intermédiaire de se sentir coupable envers le donateur même si celui-ci n'est pas affecté monétairement par l'action de l'intermédiaire.

Au-delà de ce cas spécifique du donneur dans les situations de détournement de fonds, nous nous sommes demandés, au chapitre 2, si la vulnérabilité des co-acteurs était une condition nécessaire à l'émergence de la culpabilité. L'étude d'un nouveau modulateur de la culpabilité semblait être une voie prometteuse pour mieux comprendre sa nature. En effet, l'aversion à la culpabilité s'est révélée sensible à des modulateurs tels que la communication avant le jeu entre les joueurs (*e.g.* Balafoutas and Sutter, 2017) ou le caractère raisonnable des attentes (*e.g.* Balafoutas and Fornwagner, 2017). En ce qui concerne la vulnérabilité, comme les études précédentes étaient basées sur le modèle de Battigalli and Dufwenberg (2007), dont la prémisse est que le dictateur/trusté peut se sentir coupable envers un destinataire/trusté vulnérable, elles n'ont pas pu aborder cette question. Pour combler cette lacune, nous avons conçu quatre mini-jeux de Quasi-Trust qui font systématiquement varier la vulnérabilité des co-joueurs.

Les chapitres 1 et 2 ont testé si la sphère d'influence de la culpabilité pouvait être étendue à de nouvelles situations (par exemple, lorsqu'un autre joueur n'est pas affecté financièrement, ou n'est pas vulnérable). Dans le chapitre 3, nous avons cherché à savoir si l'influence de l'aversion à la culpabilité sur le comportement pouvait être surestimée par les études expérimentales précédentes. En effet, l'aversion à la culpabilité est considérée comme une préférence pro-sociale, mais ce type de préférence tend à être remis en cause par des excuses situationnelles (par exemple, Dana et al., 2007). En fait, les paradigmes expérimentaux précédents, où l'incertitude quant aux attentes des autres est résolue lorsque les actions sont mises en œuvre, laissent peu de place à de telles excuses. Dans ce chapitre, nous avons choisi de laisser le décideur dans l'incertitude quant aux attentes des autres. Nous avons adapté le modèle d'acquisition d'informations de Spiekermann and Weiss (2016) afin de prédire comment des individus ayant des préférences dépendantes des croyances acquièrent des informations sur les attentes des autres. Enfin, une expérience nous a permis de mettre à l'épreuve nos prédictions théoriques et de déceler si certains individus acquièrent de l'information de manière stratégique afin de minimiser la tension entre leur intérêt monétaire et leur motivation dépendante des croyances.

Etat de l'art

Avancées théoriques

Dans cette section, nous décrivons plus en détail le modèle de Battigalli and Dufwenberg (2007). Nous exposons ensuite les différentes extensions de ce travail pionnier qui ont été proposées dans la littérature. Dans leur modèle d'aversion à la culpabilité, Battigalli and Dufwenberg (2007) distinguent deux concepts : la culpabilité "simple" et la culpabilité "liée aux reproches". Dans le premier cas, un joueur se soucie de savoir dans quelle mesure il déçoit les attentes d'un autre joueur, tandis que dans le second cas, un joueur

se soucie de savoir dans quelle mesure cet autre joueur peut lui reprocher d'avoir déçu ses attentes.

Nous nous concentrons sur le concept le plus courant : la culpabilité simple. [7] Dans Battigalli and Dufwenberg (2007), le décideur i ressent l'utilité de son gain matériel π_i, et la désutilité de se sentir coupable G_{ij} (Equation 2). Battigalli and Dufwenberg (2007) ont introduit la fonction de déception D_j (Equation 3) qui représente la différence, si elle est positive, entre l'espérance initiale j du co-joueur concernant son gain matériel (α_j) et le montant donné par le décideur. La culpabilité du décideur est déterminée par la part de culpabilité qui peut être attribuée à son choix s_i : $D_j(s_i, s_j) - min_{s_{\tilde{i}} D_j(s_i, s_{\tilde{i}})}$. Enfin, θ_i représente le paramètre de sensibilité à la culpabilité qui est unique à chaque individu.

$$u_i(z, s_i, \alpha_j) = \pi_i(z) - G_{ij}(z, s_i, \alpha_j) \quad (1)$$

$$\text{where } G_{ij}(z, s_i, \alpha_j) = D_j(s_i, s_j) - min_{s_{\tilde{i}}} D_j(s_i, s_{\tilde{i}}) \quad (2)$$

$$\text{and } D_j = max\{E_{s_j, \alpha_j}[\pi_j] - \pi_j, 0\} \quad (3)$$

Le modèle séminal de Battigalli and Dufwenberg (2007) a ouvert la voie à différentes lignes d'extensions. Premièrement, Khalmetski et al. (2015) ont étendu le modèle de culpabilité simple pour rendre compte de la joie des surprises positives et pas seulement de la désutilité des surprises négatives. Deuxièmement, Inderst et al. (2019) ont nuancé la définition de la culpabilité en introduisant la possibilité de blâmer le co-joueur.

Khalmetski et al. (2015) ont développé une extension innovante du modèle de Battigalli and Dufwenberg (2007) dans laquelle ils proposent que les décideurs puissent également

[7] Les articles sur la culpabilité liée aux reproches sont rares, peut-être parce que la culpabilité liée aux reproches nécessite de raisonner avec des croyances de troisième et quatrième ordre, ce qui est cognitivement exigeant (pour des exceptions, voir Charness and Dufwenberg, 2011; Beck et al., 2013).

aimer surprendre positivement les attentes de leurs co-joueurs.[8] Comme expliqué dans l'équation 5, ils ont considéré la fonction de surprise S_i où le premier terme représente l'utilité des surprises positives (lorsque $x > t_i$) et le second terme représente la désutilité des surprises négatives (lorsque $x < t_i$). α_i et β_i correspondent, respectivement, à la propension à faire des surprises positives et à éviter les surprises négatives. Enfin, le point de référence x correspond à la distribution des croyances de premier ordre, donnée par la fonction de densité h_j.[9]

$$u_i(\pi_j, h_j) = \pi_i + S_i(\pi_j, h_j) \tag{4}$$

$$\text{where } S_i(\pi_j, h_j) = \alpha_i \int_0^{p_j} (\pi_j - x) h_j(x) dx - \beta_i \int_0^{p_j} (x - \pi_j) h_j(x) dx \tag{5}$$

Dans le jeu du dictateur, cette extension prédit que les dictateurs qui ont une forte préférence pour les surprises positives verront leurs décisions de transfert corrélées négativement avec les attentes des bénéficiaires, contrairement à la culpabilité simple qui prédit une corrélation positive. En effet, lorsque les attentes du destinataire sont faibles, cela laisse plus de place au dictateur pour créer une surprise positive. Leur extension a également été appliquée dans le contexte d'un jeu de bien public par Dhami et al. (2019). Ils ont constaté que 30% des dictateurs avec des préférences dépendantes des croyances aimaient surprendre positivement leurs destinataires.

En dépit de l'élargissement de la sphère d'influence qu'ils ont proposé, Khalmetski et al. (2015) ont toujours considéré les surprises par rapport aux attentes concernant le gain

[8] Notez qu'ils ont également étendu leur modèle pour intégrer le fait que les dictateurs peuvent se soucier des inférences des destinataires sur leurs intentions (comme dans la culpabilité liée au reproche).

[9] À la différence de nombreux modèles appliqués d'aversion à la culpabilité (e.g., Beck et al., 2013), Khalmetski et al. (2015) n'ont pas pris l'espérance comme point de référence du co-joueur. Ils considèrent plutôt que le point de référence du co-joueur est stochastique : il correspond à la distribution de probabilité des résultats possibles pour le co-joueur.

propre gain du co-joueur. Dans les deux premiers chapitres de cette thèse, nous sommes revenus à ne considérer que les surprises négatives. Cependant, sur la base de jeux à trois joueurs, nous avons étendu la définition de la culpabilité à la déception des attentes concernant le gain d'un autre joueur, à savoir le gain du joueur le plus désavantagé.

Pour tester la pertinence de pouvoir blâmer un co-joueur, Inderst et al. (2019) ont examiné un jeu client-conseiller dans lequel le client peut décider entre acheter certaines informations vérifiées (*Out*) ou faire confiance au conseil du conseiller (*In*). Dans ce contexte, les auteurs ont introduit un nouveau concept, appelé culpabilité partagée, qui capture l'idée que "*l'attribution de la culpabilité pour avoir déçu les attentes est partagée entre les joueurs dont les choix ont éventuellement causé cette déception, y compris la joueur déçu lui-même*". De la même manière que Battigalli and Dufwenberg (2007) l'ont fait pour G_{ij}, Inderst et al. (2019) ont calculé τ_i et τ_j comme la part de déception qui peut être attribuée aux choix j du décideur et, respectivement, du co-joueur ; ils se réfèrent à τ_j comme l'auto-culpabilité du co-joueur. La culpabilité finale du décideur correspond à une fonction croissante de sa responsabilité (en accord avec la culpabilité simple) et décroissante de l'auto-culpabilité du co-joueur (nouveauté liée à leur formulation de culpabilité partagée).

$$U_i(\tau_i, \tau_j) = \pi_i - \theta_i G_i(\tau_i, \tau_j) \tag{6}$$

$$\text{where } \tau_i(s_i, s_j) = D_j(s_i, s_j) - min_{s_{\tilde{i}}} D_j(s_{\tilde{i}}, s_j) \tag{7}$$

$$\text{and } \tau_j(s_i, s_j) = D_j(s_i, s_j) - min_{s_{\tilde{j}}} D_j(s_i, s_{\tilde{j}}) \tag{8}$$

D'une part, le modèle de culpabilité simple prédit que plus le coût de l'information est élevé, plus les croyances de premier ordre du client conditionnelles à *In* sont élevées,

donc plus le taux de mensonge du conseiller est faible. D'autre part, en cas de culpabilité partagée, plus le coût de l'information est élevé, plus la responsabilité du client dans le choix de In est grande, plus le taux de mensonge du conseiller est élevé. Inderst et al. (2019) ont trouvé des résultats conformes aux deux prédictions étant donné l'hétérogénéité des préférences des participants.

L'étude de Inderst et al. (2019) suggère que, lorsqu'il existe un moyen d'échapper à leurs sentiments de culpabilité, les joueurs le saisissent. Le troisième chapitre de cette thèse donne suite à cette intuition en permettant aux joueurs de résoudre l'incertitude sur les attentes des autres d'une manière stratégique et intéressée.

Contributions méthodologiques

Nous allons maintenant décrire et discuter les méthodologies actuelles utilisées pour capturer l'aversion à la culpabilité. Avant de plonger dans les différentes approches pour capturer la culpabilité, il est important de noter que l'aversion à la culpabilité doit être mesurée en fonction de son anticipation plutôt que de son expression réelle. En effet, comme le souligne Miettinen and Suetens (2008), la culpabilité est une émotion contrefactuelle, c'est-à-dire une émotion induite par la pensée d'une défection qui n'a pas encore été réalisée. Si nous devions mesurer le sentiment de culpabilité après une défection réelle, plutôt que l'anticipation de la culpabilité, nous ne saisirions que la culpabilité des participants ayant la plus faible sensibilité à la culpabilité, puisque les participants ayant la plus forte sensibilité à la culpabilité auraient évité de subir le coût psychologique de la défection.

Nous discutons d'abord de deux approches différentes, avec et sans transmission de croyances, qui visent à mettre en évidence l'existence de comportements d'aversion à la

culpabilité. Cependant, l'aversion à la culpabilité n'est pas une émotion qui s'exprime en "tout-ou-rien". Par conséquent, nous examinons ensuite les études qui ont tenté d'estimer le degré de culpabilité.

Il existe deux approches principales pour capturer les comportements d'aversion à la culpabilité : avec et sans transmission de croyances. Lors de la transmission de croyances, nous distinguons trois méthodes : la méthode de *base* (demander des croyances de second ordre), la méthode de la *révélation* (divulguer des croyances de premier ordre) et la méthode du "*menu*" (conditionner leur choix à d'éventuelles croyances de premier ordre). Dans leur étude pionnière, Charness and Dufwenberg (2006) ont utilisé la méthode de *base*. Ils ont interrogé les participants sur leurs croyances de second ordre, c'est-à-dire sur ce qu'ils attendent de leurs co-joueurs. En accord avec le modèle d'aversion à la culpabilité, leurs résultats ont mis en évidence une corrélation positive entre les croyances de second ordre des participants et leurs choix quant à leur dons à leurs co-joueurs. Cependant, la corrélation observée peut avoir été causée par l'effet de faux consensus (Ross et al., 1977) : les décideurs ont tendance à croire que les autres pensent et agissent de la même manière qu'eux. Par conséquent, les décideurs peuvent avoir prédit les attentes de leurs co-joueurs en se basant sur ce qu'ils avaient l'intention de faire. Pour contrôler cet effet potentiel de faux consensus, Ellingsen et al. (2010) ont proposé un étude expérimentale dans laquelle les participants devaient indiquer leurs croyances de premier ordre sur le comportement du décideur. Ensuite, sans l'avoir dit à l'avance, ces croyances de premier ordre étaient transmises au co-joueur, afin de manipuler de manière exogène ses croyances de second ordre. Ceci correspond à la méthode de la *révélation*. Les auteurs n'ont pas trouvé de corrélation positive entre les croyances de premier ordre transmises et le comportement des joueurs et ils ont conclu que les preuves précédentes ne capturaient pas l'aversion à la culpabilité mais plutôt la magnitude de l'effet de faux consensus. Khalmetski et al. (2015) ont réconcilié ces résultats en faisant la distinction

entre les corrélations au niveau agrégé et les corrélations au niveau individuel. Comme l'a noté Tangney and Fisher (1995) et comme le permet le modèle de Battigalli and Dufwenberg (2007), le degré de sensibilité à la culpabilité peut varier d'un individu à l'autre. Par conséquent, une corrélation nulle observée au niveau agrégé peut refléter le fait que certains individus ont une corrélation positive alors que d'autres ont une corrélation négative. Pour capturer l'aversion à la culpabilité au niveau individuel, ils ont introduit la méthode du "*menu*" qui consiste à demander aux décideurs de formuler une série de choix conditionnels aux croyances possibles de premier ordre du co-joueur. Le choix effectivement mis en oeuvre pour le paiement est celui correspondant à la croyance de premier ordre réelle du co-joueur. Cette méthode a depuis été largement utilisée par d'autres expérimentateurs (par exemple, Attanasi et al., 2013 ; Hauge, 2016 ; Balafoutas and Fornwagner, 2017 ; Bellemare et al., 2017 ; Bellemare et al., 2018 ; Dhami et al., 2019). En comparant les différentes études, nous observons que la méthode de *base* et la méthode du "*menu*" ont permis aux auteurs de mettre en évidence la présence de comportements d'aversion à la culpabilité. Il est intéressant de noter que Bellemare et al. (2017) sont arrivés à la même conclusion dans une comparaison intra-étude des deux méthodes.[10] Enfin, la méthode de *base* et la méthode du "*menu*" dissimulent toutes deux certaines informations aux participants. Comme le soulignent Khalmetski et al. (2015) eux-mêmes, les "*dictateurs pourraient se méfier en apprenant, avant de faire leur choix, que les destinataires n'ont pas été informés de tous les aspects stratégiquement pertinents de la situation de décision. Cela pourrait donner l'impression que d'autres aspects de l'étude sont peut-être aussi cachés aux dictateurs.*" Pour répondre à cette critique, Khalmetski et al. (2015) ont conçu une expérience de vérification de la robustesse dans laquelle, après avoir obtenu les croyances de premier ordre des destinataires, les auteurs ont dit aux participants que les dictateurs conditionneraient leurs choix aux

[10]En fait, Bellemare et al. (2017) ont systématiquement comparé les trois méthodes précédemment citées. Ils ont également constaté que la méthode de *révélation* induit un niveau de gentillesse inconditionnelle plus élevé.

croyances de premier ordre des destinataires. Les résultats précédents ont été reproduits, suggérant ainsi que les choix des dictateurs n'étaient pas influencés par la suspicion.

Une autre approche pour capturer les comportements d'aversion à la culpabilité consiste à manipuler de manière exogène les croyances de premier ordre du co-joueur et à fournir une information complète sur cette manipulation au décideur. Cela a pour conséquence de manipuler les croyances de second ordre du décideur dans la même direction que les croyances du co-joueur. Ensuite, l'expérimentateur peut observer si la variation des croyances entraîne une variation du comportement conforme à l'aversion à la culpabilité sans avoir à transmettre les croyances réelles. Dans le cadre de cette approche, diverses expériences utilisant différents modèles ont soutenu l'hypothèse de l'aversion à la culpabilité.[11] La première utilisation de cette méthode a été proposée par Ederer and Stremitzer (2017). Dans un jeu de confiance, ils ont introduit un dispositif aléatoire qui déterminait si le second joueur pouvait décider du montant à donner au premier joueur ou si l'ordinateur décidait que le premier joueur reçoive zéro. Le premier joueur savait ex ante si ce dispositif aléatoire était fiable (forte probabilité de laisser le second joueur décider) ou non fiable (faible probabilité de laisser le second joueur décider). S'appuyant sur la même idée, Khalmetski (2016) a conçu un jeu émetteur-récepteur où les incitations matérielles de l'émetteur à mentir étaient soit faibles, soit élevées. Le récepteur connaissait la probabilité ex ante que les incitations de l'émetteur à mentir soient élevées. Enfin, dans un jeu de confiance, Inderst et al. (2019) manipulait l'option extérieure du premier joueur : plus l'option extérieure était élevée, plus le premier joueur était prêt à renoncer en choisissant In, plus les croyances de premier ordre du premier joueur étaient élevées. Dans la même veine, Balafoutas and Sutter (2017) ont révélé

[11] Toutefois, comme mentionné dans la section précédente, Ederer and Stremitzer (2017) a trouvé des preuves d'aversion à la culpabilité uniquement dans des contextes où il existait un lien prometteur direct entre le trustee et le trustor. De même, Balafoutas and Sutter (2017) ont trouvé que leur proxy pour les croyances prédisait les transferts du dictateur actuel seulement lorsqu'il y avait de la communication pré-jeu.

l'historique des transferts passés des bénéficiaires au dictateur actuel. Ces transferts passés ont servi de proxy pour rendre compte des attentes des bénéficiaires.

Les deux premiers chapitres de cette thèse ont utilisé la méthode du *menu* où les joueurs peuvent conditionner leurs choix par rapport aux attentes possibles du co-joueur, ce qui nous a permis de capturer l'aversion à la culpabilité au niveau individuel. Dans le chapitre 3, nous implementons une combinaison de la méthode du *menu* et de la manipulation de l'option extérieure dans un jeu de confiance : les joueurs pouvaient conditionner leurs choix sur les options extérieures possibles des co-joueurs, et donc sur leurs attentes possibles. L'intérêt d'une telle méthodologie réside dans le fait que les attentes des co-joueurs sont motivées par leur option extérieure et doivent être déduites par le décideur. Ces deux raisons convergent probablement pour réduire la probabilité que le décideur minimise ces attentes.

Au-delà de la mise en évidence de l'existence de comportements d'aversion à la culpabilité, les chercheurs ont tenté d'estimer le paramètre de sensibilité à la culpabilité (θ_i dans le modèle de culpabilité simple). [12] Là encore, diverses méthodologies ont été utilisées : (i) estimation structurale, (ii) inférences des prédictions d'équilibre, (iii) inférences des limites d'information ou (iv) questionnaires hypothétiques. Bellemare et al. (2011) ont utilisé un modèle structural pour estimer la volonté de payer pour éviter de décevoir les

[12] Si le dépassement de la vision dichotomique de la mise en évidence (ou non) de l'aversion à la culpabilité est une approche prometteuse, elle a également été développée par des psychologues à l'aide de questionnaires de propension à la culpabilité : le Guilt and Shame Proneness questionnaire développé par Cohen et al. (2011), le Test of Self-Conscious Affect développé par Tangney et al. (1989) ou une question unique proposée par Moulton et al. (1966). Cependant, les études testant la cohérence entre la sensibilité à la culpabilité mesurée par les économistes et la tendance à la culpabilité mesurée par questionnaire ne sont pas concluantes. Bellemare et al. (2019) a trouvé une corrélation positive et forte entre la tendance à la culpabilité évaluée par le Test of Self-Conscious Affect et le paramètre estimé de la sensibilité à la culpabilité, tandis que Peeters and Vorsatz (2021) a rapporté une absence de corrélation entre le paramètre de culpabilité et la tendance à la culpabilité évaluée par le questionnaire Guilt and Shame Proneness.

attentes du co-joueur dans un jeu de dictateur.[13] Bellemare et al. (2018) ont mis à jour leur propre estimation en permettant à la sensibilité à la culpabilité de dépendre de la taille des enjeux monétaires, ce qui expliquait 60% des comportements des dictateurs dans leur expérience. Deuxièmement, Peeters and Vorsatz (2021) ont calculé l'ensemble de tous les équilibres dans un dilemme du prisonnier (c'est-à-dire le taux de coopération) en fonction du paramètre de culpabilité. Ils ont ensuite estimé le paramètre de culpabilité observé dans la population à partir des taux de coopération expérimentaux. Dans différents jeux de participation, Patel and Smith (2019) ont utilisé une technique similaire pour calculer le paramètre de sensibilité à la culpabilité en se basant sur les prédictions théoriques des équilibres symétriques mixtes. Troisièmement, Bellemare et al. (2019) ont proposé d'inférer des limites d'information sur le paramètre de culpabilité sans données ni hypothèses sur les croyances.[14] Enfin, Attanasi et al. (2016) et Peeters and Vorsatz (2021) ont proposé des méthodes hypothétiques pour obtenir le paramètre de sensibilité à la culpabilité. Dans le cadre d'un mini-jeu de confiance, Attanasi et al. (2016) ont demandé aux seconds joueurs d'envisager une situation dans laquelle le premier joueur a choisi *In* et où ils ont choisi de ne rien renvoyer. Ensuite, ils ont demandé combien les seconds joueurs étaient prêts à rembourser au premier joueur, pour chaque croyance de premier ordre possible du premier joueur. Dans un dilemme du prisonnier, Peeters and Vorsatz (2021) ont demandé aux participants quel était le montant minimum pour lequel ils seraient indifférents entre coopérer et faire défection.

[13] Le plan expérimental de Bellemare et al. (2011) impliquait un traitement où les participants étaient invités à déclarer leurs croyances de second ordre et un traitement où les participants étaient informés des croyances de premier ordre de leur co-joueur. Le paramètre estimé de sensibilité à la culpabilité était significativement plus élevé dans le premier traitement que dans le second, ce qui suggère la présence d'un effet de faux consensus.

[14] Malheureusement, leur analyse a donné des estimations invraisemblablement élevées de l'aversion à la culpabilité, qui étaient très probablement dues à la très faible proportion de joueurs ayant cette préférence dans leur expérience.

Dans les chapitres 1 et 2, nous avons suivi Bellemare et al. (2011) en utilisant un modèle structurel pour estimer la sensibilité à la culpabilité des décideurs. Ceci était essentiel dans notre approche qui visait à diversifier les contextes dans lesquels nous pouvons observer l'aversion à la culpabilité. En effet, en procédant ainsi, nous étions en mesure, non seulement d'évaluer l'existence de la culpabilité dans différents contextes, mais aussi de mettre en évidence si son intensité variait selon les situations.

Aperçu de la thèse

La littérature examinée jusqu'à présent a montré que l'aversion à la culpabilité existe bel et bien. Cependant, nous ne savons pas grand-chose des facteurs qui augmentent ou diminuent sa prévalence ou son impact. L'étude de certains de ces facteurs est l'objet de cette thèse. Dans le chapitre 1, dans le cadre de la situation particulière de détournement de fonds, nous avons cherché à savoir si un individu peut se sentir coupable envers un joueur qui n'est pas affecté monétairement. Dans le chapitre 2, nous avons étendu cette question à de multiples scénarios et avons fait varier de manière systématique la vulnérabilité des co-joueurs. Dans le chapitre 3, nous avons cherché à savoir si l'impact de l'aversion à la culpabilité sur le comportement peut être "illusoire" en étudiant le comportement des individus lorsqu'ils ont la possibilité d'acquérir stratégiquement des informations sur les croyances des autres.

Aversion à la culpabilité et détournement de fonds

Dans le chapitre 1, nous avons examiné une situation dans laquelle un donateur peut envoyer de l'argent à un bénéficiaire. Comme dans le cas des dons de charité ou des transferts gouvernementaux vers les pays en développement, cet argent doit être transmis par un intermédiaire qui peut en détourner une partie. Alors que les recherches précédentes ont exclusivement considéré la culpabilité de l'intermédiaire envers le bénéficiaire

(comme dans les jeux de dictateur), nous avons envisagé qu'un intermédiaire puisse également éprouver de la culpabilité envers le donateur qui a donné cette somme avec l'intention de la reverser au bénéficiaire. Nous avions pour objectif de vérifier si la direction de la culpabilité (envers le bénéficiaire ou envers le donateur) affectait l'intensité ou la prévalence de l'aversion à la culpabilité chez les intermédiaires.

Nous avons conçu un nouveau jeu à trois joueurs, le mini-jeu de détournement de fonds. Dans ce jeu, un donateur envoie un don à un bénéficiaire, mais ce don doit être transféré par un intermédiaire qui peut détourner une partie de ce don pour augmenter son propre gain matériel. Le donateur forme des attentes sur la quantité du don que l'intermédiaire transférera au bénéficiaire, et le bénéficiaire forme des attentes sur la quantité qu'il/elle recevra. Par conséquent, l'intermédiaire peut décevoir deux types d'attentes. Pour capturer l'aversion de l'intermédiaire à décevoir les attentes, c'est-à-dire sa culpabilité, nous avons permis à l'intermédiaire de conditionner sa décision de transfert aux attentes du donateur (traitement du donateur) ou du bénéficiaire (traitement du bénéficiaire). Cette manipulation a été effectuée entre les sujets et nous a permis de capturer l'aversion à la culpabilité au niveau individuel : les intermédiaires qui ont augmenté leurs transferts en fonction des attentes de leur co-joueur ont été classés comme averses à la culpabilité. De plus, nous avons fait varier pour un même sujet le pourcentage du don qui pouvait être détourné (80% dans la condition Haute et 60% dans la condition Basse) afin de tester dans quelle mesure l'intensité du détournement potentiel affectait les croyances.

En ce qui concerne la culpabilité envers le bénéficiaire, nous nous sommes appuyés sur la définition de la culpabilité donnée par Battigalli and Dufwenberg (2007), à savoir la désutilité liée au fait de décevoir les attentes du bénéficiaire concernant son propre gain. En ce qui concerne la culpabilité envers le donneur, nous avons étendu théoriquement le modèle Battigalli and Dufwenberg (2007). Plutôt que de ne pas décevoir les attentes du

donneur quant à son propre gain (qui n'est pas affecté par la décision de détourner des fonds), un intermédiaire averse à la culpabilité envers le donneur n'aime pas décevoir les attentes du donneur quant au gain matériel d'un autre joueur, c'est-à-dire le bénéficiaire.

Comme attendu, nous avons constaté que l'aversion à la culpabilité réduit le détournement de fonds chez les intermédiaires. De plus, nos résultats expérimentaux ont indiqué que la proportion d'intermédiaires qui éprouvent de la culpabilité est similaire, que la culpabilité soit envers le bénéficiaire ou envers le donateur (en moyenne, 25%). L'absence de différence entre les traitements est confirmée lorsqu'on examine l'intensité de leurs aversions (en moyenne, un intermédiaire est prêt à payer 0,37 écu pour ne pas décevoir un autre joueur de 1 écu). Ce résultat est frappant car le détournement de fonds affecte les gains du bénéficiaire mais pas ceux du donateur. Il montre que les mécanismes en jeu dans l'aversion à la culpabilité s'étendent à des situations où les décisions n'ont pas de conséquences monétaires directes. Ces résultats sont valables quel que soit le pourcentage du don qui pouvait être détourné.

Aversion à la culpabilité et vulnérabilité

Au chapitre 1, nous avons démontré qu'un individu (l'intermédiaire) peut se sentir coupable même envers une personne qui n'est pas affectée financièrement (le donateur). Sur la base de cette constatation, nous avons ensuite exploré le rôle de la vulnérabilité dans le déclenchement de la culpabilité d'une personne.

Sur la base des résultats du chapitre précédent, nous avons distingué deux types de vulnérabilité dans le chapitre 2. D'une part, un individu peut être vulnérable "ex-post" si son gain final dépend de l'action du décideur. D'autre part, un individu peut être vulnérable "ex-ante" si sa dotation initiale peut être confiée au décideur. Dans ce chapitre,

nous avons cherché à évaluer si l'aversion à la culpabilité est modulée par les différentes combinaisons des deux types de vulnérabilité du joueur envers lequel le décideur peut se sentir coupable. De plus, nous avons voulu tester si les deux types de vulnérabilité sont complémentaires ou substitutifs dans leur impact sur la culpabilité.

Pour répondre à ces questions, nous avons conçu une expérience en laboratoire avec quatre jeux en deux étapes avec deux joueurs actifs (A et B) et un joueur passif (C). Dans chaque jeu, le second joueur (B) peut se voir confier par le premier joueur (A) une somme d'argent. Cet argent provient de la dotation du joueur A ou C, selon le jeu - vulnérabilité ex-ante du joueur A ou C. Ensuite, le joueur B peut redistribuer cet argent entre lui et un autre joueur (A ou C, selon le jeu) - vulnérabilité ex-post du joueur A ou C. Ces quatre mini-jeux faisaient varier systématiquement la vulnérabilité des joueurs A et C. Ils étaient joués par un même sujet dans un ordre aléatoire. De plus, nous avons manipulé entre les sujets le fait que le comportement du joueur B soit élicité conditionnellement aux croyances de premier ordre du joueur A ou C. Ce faisant, nous avons manipulé le fait que l'aversion à la culpabilité du joueur B soit déclenchée envers un joueur dont les intentions sont observables (joueur A, actif) ou non (joueur C, passif).

Sur le plan théorique, nous suivons le modèle exposé au chapitre 1, qui soutient que la culpabilité est activée même lorsque les croyances du joueur déçu ne concernent pas son gain matériel mais le gain d'un troisième joueur. Nous sommes allés un peu plus loin en permettant à l'aversion à la culpabilité du joueur B d'être déclenchée même lorsque l'intention du joueur A était médiée par les croyances de premier ordre d'un autre joueur, à savoir le joueur passif. En d'autres termes, notre modèle théorique prédit que l'aversion à la culpabilité ne dépend ni du jeu, ni du statut du joueur déçu (actif *vs* passif).

Nous avons trouvé des manifestations d'aversion à la culpabilité dans les quatre jeux. Ceci a révélé la pertinence de l'aversion à la culpabilité dans des jeux où elle n'avait jamais été testée auparavant. En accord avec nos prédictions théoriques, nos résultats n'ont montré aucune différence significative ni dans la proportion de joueurs B ayant une aversion à la culpabilité ni dans leur intensité de culpabilité. Certains joueurs B ont montré un comportement d'aversion à la culpabilité même envers des joueurs qui n'étaient vulnérables dans aucunes des deux dimensions. Enfin, le fait d'observer ou non l'intention des co-joueurs ne semble pas moduler l'aversion à la culpabilité du décideur.

Aversion à la culpabilité et acquisition d'informations

Le chapitre 2 a étendu la sphère d'influence de l'aversion à la culpabilité en révélant son existence dans une variété de situations où elle n'avait jamais été testée auparavant. Le chapitre 3 a abordé une autre limite des études précédentes et a remis en question la force de cette préférence en utilisant un design où les joueurs peuvent auto-sélectionner les informations sur les attentes des autres, ce qui nous a permis d'explorer les stratégies utilisées par les individus pour éviter de se sentir coupables tout en se comportant de manière égoïste.

De nombreux éléments montrent que les individus se soucient du bien-être d'autrui (pour une étude, voir Cooper, 2009). Pourtant, il a été démontré que ces préférences apparemment prosociales s'estompent en présence d'une incertitude sur la relation entre ses propres actions et leurs conséquences (par exemple, Dana et al., 2007). En revanche, on sait peu de choses sur la robustesse des préférences lorsque l'incertitude concerne les attentes des autres. Dans ce ce chapitre, nous avons abordé cette dernière question en examinant si les individus ayant des préférences dépendantes des croyances biaisaient leur stratégie d'acquisition d'informations afin de minimiser la tension entre leur intérêt

monétaire et leur motivation dépendante des croyances.

Nous avons adapté le modèle d'acquisition d'informations de Spiekermann and Weiss (2016) pour étudier si les agents averses à la culpabilité et réciproques acquièrent stratégiquement des informations sur les attentes des autres. Cette approche permet de distinguer les préférences objectives des préférences subjectives. Appliquées aux domaine des préférences dépendantes des croyances, les préférences objectives impliquent que les agents maximisent leur utilité en se conformant aux attentes réelles des autres alors que les préférences subjectives permettent aux agents de maximiser leur utilité en se conformant à la croyance qu'ils ont sur les attentes des autres. Notre modèle prédit que les agents ayant des préférences objectives dépendantes des croyances recherchent toujours plus d'informations, quelle que soit leur motivation dépendante des croyances (aversion à la culpabilité ou réciprocité), tandis que les agents ayant des préférences subjectives dépendantes des croyances recherchent stratégiquement les informations qui minimisent la tension entre leur intérêt monétaire et leur motivation dépendante des croyances.

Nous avons testé nos prédictions dans une expérience en ligne. Nous avons conçu un jeu de confiance modifié dans lequel nous avons manipulé l'option extérieure du premier joueur afin d'influencer les croyances de premier ordre du premier joueur et, par anticipation, les croyances de second ordre du second joueur. Dans les faits, nous avons créé une variation exogène des croyances pour 61,25 % des seconds joueurs de notre échantillon. Nous avons ensuite pu révéler les préférences des seconds joueurs en leur demandant de déclarer leurs choix de retour conditionnellement à la connaissance de l'option extérieure du premier joueur. Nous avons constaté que 52,04% des seconds joueurs avaient des préférences indépendantes des croyances, 43,88% étaient averses à

la culpabilité et 4,08% étaient réciproques. Enfin, les seconds joueurs ont eu l'occasion inattendue d'acquérir des informations sur l'option extérieure du premier joueur. Il est important de noter que le choix mis en œuvre par les seconds joueurs dépendait des informations dont ils disposaient : certitude que l'option extérieure est faible, certitude que l'option extérieure est élevée ou incertitude quant à l'option extérieure.

Nous avons constaté que la majorité des seconds joueurs ayant des préférences dépendantes des croyances présentaient une stratégie d'acquisition d'informations conforme aux préférences subjectives : 60,47% des seconds joueurs ayant une aversion à la culpabilité ont cherché à obtenir un signal bas uniquement. Les analyses de régression et les réponses au questionnaire post-expérimental suggèrent que ce choix a été fait pour maximiser son propre gain. Symétriquement, le seul second joueur réciproque de notre échantillon a cherché un signal élevé uniquement. Enfin, il convient de mentionner qu'une fraction non négligeable de notre échantillon a acquis des informations selon un modèle compatible avec les préférences objectives dépendant des croyances (20,93%). Nos résultats suggèrent que l'impact positif des préférences dépendantes des croyances sur les choix pro-sociaux dépend de la (in)certitude vis-à-vis des croyances des autres joueurs.

Pour conclure, toute l'ambition de cette thèse était de mieux cerner la sphère d'influence de l'aversion à la culpabilité. Nos travaux ont révélé pour la toute première fois que l'aversion à la culpabilité peut également être déclenchée lorsque la personne envers laquelle on se sent coupable n'est pas affectée monétairement. Ce résultat, obtenu dans le cadre d'un jeu de détournement de fonds, a ensuite été étendu à de nouveaux contextes où l'aversion à la culpabilité était systématiquement observée envers les joueurs indépendamment de leur vulnérabilité. Enfin, bien que l'aversion à la culpabilité semble se généraliser à une variété de situations, nous avons démontré que sa robustesse peut

être remise en question dans des situations où les décideurs ont la possibilité d'éviter la tension entre leurs incitations monétaires et leurs préoccupations liées aux croyances.

Discussion de la thèse

L'objectif de cette thèse était de questionner la sphère d'influence de l'aversion à la culpabilité sur le comportement. Nous avons montré que son champ d'influence est plus large qu'initialement soupçonné mais qu'il s'agit d'un phénomène moins robuste que ce que la littérature précédente suggérait. D'une part, nous avons proposé un modèle et démontré expérimentalement que la portée de l'influence de la culpabilité s'étend vers les joueurs qui ne sont pas affectés monétairement par l'action du décideur. D'autre part, nous avons révélé que les individus développent des stratégies pour acquérir des informations sur les attentes des autres qui leur permettent de mettre en œuvre l'option la moins pro-sociale sans subir le coût psychologique de l'aversion à la culpabilité.

Dans le chapitre 1, nous avons examiné une situation dans laquelle un donateur peut envoyer de l'argent à un bénéficiaire. Comme dans le cas des dons de charité ou des transferts gouvernementaux vers les pays en développement, cet argent doit être transmis par un intermédiaire qui peut en détourner une partie. Nous avons testé si la direction de la culpabilité (envers le bénéficiaire ou envers le donateur) affecte l'intensité ou la prévalence de l'aversion à la culpabilité chez les intermédiaires. Nos résultats expérimentaux ont indiqué que la proportion d'intermédiaires qui éprouvent de la culpabilité était similaire, que la culpabilité soit envers le bénéficiaire ou envers le donateur. L'absence de différence entre les traitements a été confirmée lorsqu'on a examiné l'intensité de leurs aversions. Ce résultat est frappant car le détournement de fonds affecte les gains du bénéficiaire mais pas ceux du donateur. Cela montre que les mécanismes en jeu dans l'aversion à la culpabilité s'étendent à des situations où les décisions n'ont pas de

conséquences monétaires directes.

Dans ce chapitre, nous avons mesuré l'aversion à la culpabilité envers le donneur et envers le receveur dans deux traitements distincts. Étant donné que nous sommes les premiers à documenter l'existence de la culpabilité envers le donneur, la littérature est agnostique quant à l'effet conjoint de ces deux types d'attentes. Par conséquent, une extension importante de cette étude consisterait à tester un traitement dans lequel les intermédiaires seraient informés à la fois des attentes des donateurs et des bénéficiaires : cela permettrait de tester si un effet l'emporte sur l'autre, ou si leurs effets sont cumulatifs. Cela conduirait toutefois à une étude complexe. En outre, une autre extension de la présente étude incluant plusieurs tours de jeu, au lieu d'un jeu à un coup, pourrait permettre de tester l'hypothèse de Balafoutas (2011) d'un cercle vicieux de normes corrompues. Si les donateurs ou les bénéficiaires s'attendent à un niveau élevé de détournement de fonds dans un groupe, les intermédiaires peuvent détourner des fonds sans se sentir coupables, ce qui augmente les attentes de détournement de fonds.

Dans le chapitre 2, nous avons évalué si l'aversion à la culpabilité est modérée par la vulnérabilité du joueur envers lequel le décideur peut se sentir coupable. Pour ce faire, nous avons conçu quatre nouveaux mini-jeux de Quasi-Trust où nous avons fait varier systématiquement la vulnérabilité des co-joueurs. Nous avons constaté que ni la proportion de seconds joueurs averses à la culpabilité ni l'intensité de leur aversion à la culpabilité ne différaient de manière significative entre les quatre jeux (c'est-à-dire entre les quatre combinaisons de vulnérabilité), et entre les deux traitements (c'est-à-dire entre la culpabilité envers un joueur actif *vs.* un joueur passif). En particulier, les seconds joueurs présentent un comportement d'aversion à la culpabilité, même à l'égard des croyances des joueurs qui ne sont pas du tout vulnérables. Il est intéressant de noter que cela

confirme la pertinence de l'aversion à la culpabilité dans les situations d'investissement et révèle son importance dans les situations de don ou d'exploitation.

En résumé, nous avons montré que la vulnérabilité des co-joueurs n'affecte pas l'aversion à la culpabilité. L'insensibilité de l'aversion à la culpabilité des seconds joueurs aux manipulations des gains et des intentions des co-joueurs pourrait cependant être interprétée comme un signe de confusion de la part de nos sujets (c'est-à-dire que les sujets n'ont pas compris les différents jeux). Pourtant, le comportement des premiers joueurs plaide contre cette interprétation. Nous avons constaté que leur comportement dépendait du jeu, conformément à notre modèle d'altruisme lexicographique. Par ailleurs, l'affichage des choix des joueurs B pourrait avoir réduit l'impact de la vulnérabilité des co-joueurs. En effet, il était demandé aux participants de rendre leur choix conditionnel à quatre niveaux d'attentes de l'autre joueur. Cette contextualisation des choix, traditionnelle lorsqu'on teste des préférences basées sur des croyances, a potentiellement neutralisé l'information fournie lors de l'introduction du jeu qui était censée déclencher une réaction basée sur la vulnérabilité de l'autre joueur. Cette explication alternative de nos résultats pourrait être testée en informant des attentes des co-joueurs au début du jeu et en demandant aux seconds joueurs de conditionner leurs choix aux différentes manipulations de la vulnérabilité de leurs co-joueurs.

Dans le chapitre 3, nous avons cherché à savoir si les individus ayant des préférences dépendantes des croyances biaisent leur stratégie d'acquisition d'informations afin de minimiser la tension entre leur intérêt monétaire et leur motivation dépendante des croyances. Nous avons testé nos prédictions dans une expérience en ligne où nous avons donné l'opportunité aux seconds joueurs d'acquérir des informations sur l'option extérieure des premiers joueurs, et ce faisant sur leurs attentes. Nos résultats suggèrent que l'influence de l'aversion à la culpabilité dépend de la (in)certitude sur les attentes des

co-joueurs. Notre principale contribution est de montrer que les individus orientent leur stratégie d'acquisition d'informations vers des signaux bénéfiques pour eux afin d'éviter de payer le coût monétaire que suivre leur conscience impliquerait, c'est-à-dire en faisant le choix qui correspond à l'état réel du monde. Dans la littérature traditionnelle sur les préférences dépendantes des croyances, les attentes du premier joueur sont parfaitement observables par le second joueur au moment où ses actions sont réalisées. Nos résultats suggèrent que cette littérature capture une estimation supérieure de l'impact positif des préférences dépendantes des croyances sur les choix pro-sociaux.

Il est à noter que notre modèle théorique et notre conception expérimentale considèrent tous deux le cas d'un modélisation "grossière" des croyances, c'est-à-dire que la croyance sur l'état du monde est une fonction à échelon. Si nous devions relâcher cette caractéristique et permettre une modélisation linéaire des croyances, le choix optimal d'un second joueur avec des préférences "illusoires" dépendantes des croyances serait d'éviter toute information, un choix qui n'est pas possible dans notre expérience. Par conséquent, une extension naturelle serait de considérer une modélisation linéaire des croyances dans un dispositif expérimental qui permettrait d'ignorer toute information. Cependant, cette extension nécessite que les participants soient capables de mettre à jour leurs croyances de manière bayésienne, ce qui s'est avéré assez difficile (par exemple, Grether, 1980 ; Belot et al., 2012). En outre, on peut se demander si les stratégies d'acquisition de l'information sont influencées par "l'option par défaut". Dans la présente étude, les participants devaient choisir les informations qu'ils souhaitaient acquérir. Que se passerait-il si, par défaut, toutes les informations étaient sélectionnées, et que les participants devaient désélectionner les informations qu'ils ne voulaient pas connaitre ? Les résultats de Grossman and Van Der Weele (2017) suggèrent que ce dispositif experimental alternatif pourrait conduire à une acquisition moins stratégique des informations.

En termes d'application aux politiques publiques, les résultats de cette thèse suggèrent d'infléchir la communication dans les politiques anti-corruption. Les campagnes publiques d'information (Reinikka and Svensson, 2004) se concentrent généralement sur les attentes des bénéficiaires potentiels. Les résultats de cette thèse incitent à faire connaitre non seulement les attentes élevées des bénéficiaires mais aussi celles des donateurs afin de limiter les détournements de fonds par les intermédiaires. Plus généralement, si nos données renforcent l'idée de la pertinence de campagnes visant à modifier le comportement des citoyens en s'appuyant sur leur aversion à la culpabilité, elles suggèrent de veiller à ne pas rendre les informations sur les attentes des autres seulement disponibles mais à les rendre incontournables. En effet, nos données indiquent qu'en cas d'incertitude concernant les croyances d'autrui, une majorité d'individus averses à la culpabilité cherchent des informations bénéfiques pour eux. Ceci les conduit finalement à choisir l'action la plus rentable sans compromettre leur motivation dépendante des croyances.

Jusqu'à présent, les économistes se sont concentrés sur l'anticipation de la culpabilité qui, si elle est aversive, motive les comportements prosociaux. Pourtant, les psychologues ont mis en évidence que l'expérience de la culpabilité peut également encourager un comportement prosocial (Baumeister et al., 1994). En effet, l'expérience de la culpabilité peut motiver un comportement réparateur (par exemple, Ketelaar and Tung Au, 2003). La prise en compte de cet aspect de la culpabilité pourrait enrichir la littérature économique sur les excuses qui s'est concentrée sur les excuses déclenchées par le préjudice subi, plutôt que par la déception des attentes (par exemple, Abeler et al., 2010). En outre, l'affichage de la culpabilité peut apaiser les victimes ou les spectateurs.

Les expériences des chapitres 1 et 2 ont été menées en laboratoire tandis que, dans le chapitre 3, nous avons mis en œuvre une expérience en ligne sur Amazon Mechanical Turk en raison de la pandémie mondiale de Covid-19. Cette dernière situation nous a permis de tester l'existence de comportements d'aversion à la culpabilité dans un contexte qui augmente la distance sociale. Contrairement aux travaux qui suggèrent que la distance sociale limite l'aversion à la culpabilité (Morell, 2019), nous avons observé ce type de comportements chez environ 40% des participants. Cela suggère que l'aversion à la culpabilité se manifeste dans un plus large éventail de situations que celles habituellement étudiées (voir également Bellemare et al., 2011 pour la seule autre étude en ligne sur l'aversion à la culpabilité). Ces résultats ouvrent la voie à de nouvelles pistes de recherche pour étudier la culpabilité avec des échantillons beaucoup plus importants.

Tant pour les expériences en laboratoire que pour les expériences en ligne, on peut s'interroger sur la validité externe de leurs résultats. Nous ne connaissons qu'un seul article qui étudie comment l'aversion à la culpabilité suscitée dans une expérience en laboratoire peut prédire le comportement sur le terrain (Shoji, 2020). Au Bangladesh, cet auteur a montré que les personnes ayant une plus grande sensibilité à la culpabilité ont une plus grande accessibilité au crédit et une plus grande solvabilité. En outre, les individus souffrent moins de crimes contre la propriété dans les villages où la sensibilité à la culpabilité est plus élevée. Ces résultats suggèrent que l'aversion à la culpabilité mesurée en laboratoire peut effectivement expliquer les comportements sur le terrain. Ils renforcent notre confiance dans les implications politiques potentielles de nos expériences de laboratoire actuelles. Cependant, une étape supplémentaire serait nécessaire, qui impliquerait de mesurer l'aversion à la culpabilité directement à travers les comportements sur le terrain. Le défi majeur de cette transposition sera de mesurer les croyances à l'origine de cette aversion car elles sont difficilement observables sur le terrain.

Bibliography

Abeler, J., Calaki, J., Andree, K., and Basek, C. (2010). The power of apology. *Economics Letters*, 107(2):233–235.

Attanasi, G., Battigalli, P., and Manzoni, E. (2016). Incomplete-information models of guilt aversion in the trust game. *Management Science*, 62(3):648–667.

Attanasi, G., Battigalli, P., and Nagel, R. (2013). Disclosure of belief-dependent preferences in a trust game. Technical report, No. 506, IGIER (Innocenzo Gasparini Institute for Economic Research), Bocconi University.

Azar, O. H. (2019). The influence of psychological game theory. *Journal of Economic Behavior & Organization*, 167:445–453.

Balafoutas, L. (2011). Public beliefs and corruption in a repeated psychological game. *Journal of Economic Behavior & Organization*, 78(1-2):51–59.

Balafoutas, L. and Fornwagner, H. (2017). The limits of guilt. *Journal of the Economic Science Association*, 3(2):137–148.

Balafoutas, L. and Sutter, M. (2017). On the nature of guilt aversion: Insights from a new methodology in the dictator game. *Journal of Behavioral and Experimental Finance*, 13:9–15.

Battigalli, P., Charness, G., and Dufwenberg, M. (2013). Deception: The role of guilt. *Journal of Economic Behavior & Organization*, 93:227–232.

Battigalli, P. and Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, 97(2):170–176.

Battigalli, P. and Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory*, 144(1):1–35.

Battigalli, P. and Dufwenberg, M. (2020). Belief-dependent motivations and psychological game theory.

Battigalli, P., Dufwenberg, M., and Smith, A. (2019). Frustration, aggression, and anger in leader-follower games. *Games and Economic Behavior*, 117:15–39.

Baumeister, R. F., Reis, H. T., and Delespaul, P. A. (1995). Subjective and experiential correlates of guilt in daily life. *Personality and Social Psychology Bulletin*, 21(12):1256–1268.

Baumeister, R. F., Stillwell, A. M., and Heatherton, T. F. (1994). Guilt: an interpersonal approach. *Psychological bulletin*, 115(2):243.

Beck, A., Kerschbamer, R., Qiu, J., and Sutter, M. (2013). Shaping beliefs in experimental markets for expert services: Guilt aversion and the impact of promises and money-burning options. *Games and Economic Behavior*, 81:145–164.

Bell, D. E. (1985). Reply—putting a premium on regret. *Management Science*, 31(1):117–122.

Bellemare, C., Sebald, A., and Strobel, M. (2011). Measuring the willingness to pay to avoid guilt: estimation using equilibrium and stated belief models. *Journal of Applied Econometrics*, 26(3):437–453.

Bellemare, C., Sebald, A., and Suetens, S. (2017). A note on testing guilt aversion. *Games and Economic Behavior*, 102:233–239.

Bellemare, C., Sebald, A., and Suetens, S. (2018). Heterogeneous guilt sensitivities and incentive effects. *Experimental Economics*, 21(2):316–336.

Bellemare, C., Sebald, A., and Suetens, S. (2019). Guilt aversion in economics and psychology. *Journal of Economic Psychology*, 73:52–59.

Belot, M., Bhaskar, V., and Van De Ven, J. (2012). Can observers predict trustworthiness? *Review of Economics and Statistics*, 94(1):246–259.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American economic review*, 96(5):1652–1678.

Caplin, A. and Leahy, J. (2004). The social discount rate. *Journal of political Economy*, 112(6):1257–1268.

Cartwright, E. (2019a). Guilt aversion and reciprocity in the performance-enhancing drug game. *Journal of Sports Economics*, 20(4):535–555.

Cartwright, E. (2019b). A survey of belief-based guilt aversion in trust and dictator games. *Journal of Economic Behavior & Organization*, 167:430–444.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6):1579–1601.

Charness, G. and Dufwenberg, M. (2011). Participation. *American Economic Review*, 101(4):1211–37.

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

Ciccarone, G., Di Bartolomeo, G., and Papa, S. (2020). The rationale of in-group favoritism: An experimental test of three explanations. *Games and Economic Behavior*, 124:554–568.

Cohen, T. R., Wolf, S. T., Panter, A. T., and Insko, C. A. (2011). Introducing the gasp scale: a new measure of guilt and shame proneness. *Journal of personality and social psychology*, 100(5):947.

Conconi, P., DeRemer, D. R., Kirchsteiger, G., Trimarchi, L., and Zanardi, M. (2017). Suspiciously timed trade disputes. *Journal of International Economics*, 105:57–76.

Cooper, D. (2009). Other regarding preferences: a survey of experimental results in j. kagel & a. roth. *The handbook of experimental economics*, 2.

d'Adda, G., Drouvelis, M., and Nosenzo, D. (2016). Norm elicitation in within-subject designs: Testing for order effects. *Journal of Behavioral and Experimental Economics*, 62:1–7.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.

Danilov, A., Khalmetski, K., and Sliwka, D. (2019). Descriptive norms and guilt aversion. *Mimeo*.

DellaVigna, S., Lindner, A., Reizer, B., and Schmieder, J. F. (2017). Reference-dependent job search: Evidence from hungary. *The Quarterly Journal of Economics*, 132(4):1969–2018.

Dhami, S., Wei, M., and al Nowaihi, A. (2019). Public goods games and psychological utility: Theory and evidence. *Journal of Economic Behavior & Organization*, 167:361–390.

Di Bartolomeo, G., Dufwenberg, M., Papa, S., and Passarelli, F. (2018). Promises, expectations & causation. *Games and Economic Behavior*.

Dollard, J., Miller, N. E., Doob, L. W., Mowrer, O. H., and Sears, R. R. (1939). Frustration and aggression.

Dufwenberg, M. (2002). Marital investments, time consistency and emotions. *Journal of Economic Behavior & Organization*, 48(1):57–69.

Dufwenberg, M. and Dufwenberg, M. A. (2018). Lies in disguise–a theoretical analysis of cheating. *Journal of Economic Theory*, 175:248–264.

Dufwenberg, M., Gächter, S., and Hennig-Schmidt, H. (2011). The framing of games and the psychology of play. *Games and Economic Behavior*, 73(2):459–478.

Dufwenberg, M. and Gneezy, U. (2000). Measuring beliefs in an experimental lost wallet game. *Games and economic Behavior*, 30(2):163–182.

Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and economic behavior*, 47(2):268–298.

Dufwenberg, M. and Kirchsteiger, G. (2019). Modelling kindness. *Journal of Economic Behavior & Organization*, 167:228–234.

Dufwenberg, M. and Nordblom, K. (2018). Tax evasion with a conscience.

Easterlin, R. A. (1995). Will raising the incomes of all increase the happiness of all? *Journal of Economic Behavior & Organization*, 27(1):35–47.

Ederer, F. and Stremitzer, A. (2017). Promises and expectations. *Games and Economic Behavior*, 106:161–178.

Ellingsen, T., Johannesson, M., Tjøtta, S., and Torsvik, G. (2010). Testing guilt aversion. *Games and Economic Behavior*, 68(1):95–107.

Elster, J. (1998). Emotions and economic theory. *Journal of economic literature*, 36(1):47–74.

Farber, H. S. (2005). Is tomorrow another day? the labor supply of new york city cabdrivers. *Journal of political Economy*, 113(1):46–82.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3):817–868.

Geanakoplos, J., Pearce, D., and Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and economic Behavior*, 1(1):60–79.

Ghidoni, R. and Ploner, M. (2020). When do the expectations of others matter? experimental evidence on distributional justice and guilt aversion. *Theory and Decision*, pages 1–46.

Gill, D. and Prowse, V. (2012). A structural analysis of disappointment aversion in a real effort competition. *American Economic Review*, 102(1):469–503.

Grether, D. M. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly journal of economics*, 95(3):537–557.

Grossman, Z. and Van Der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association*, 15(1):173–217.

Guala, F., Mittone, L., and Ploner, M. (2013). Group membership, team preferences, and expectations. *Journal of Economic Behavior & Organization*, 86:183–190.

Güth, W., Ploner, M., and Regner, T. (2009). Determinants of in-group bias: Is group affiliation mediated by guilt-aversion? *Journal of Economic Psychology*, 30(5):814–827.

Hauge, K. E. (2016). Generosity and guilt: The role of beliefs and moral standards of others. *Journal of Economic Psychology*, 54:35–43.

Heidhues, P. and Kőszegi, B. (2008). Competition and price variation when consumers are loss averse. *American Economic Review*, 98(4):1245–68.

Inderst, R., Khalmetski, K., and Ockenfels, A. (2019). Sharing guilt: How better access to information may backfire. *Management Science*, 65(7):3322–3336.

Ketelaar, T. and Tung Au, W. (2003). The effects of feelings of guilt on the behaviour of uncooperative individuals in repeated social bargaining games: An affect-as-information interpretation of the role of emotion in social interaction. *Cognition and emotion*, 17(3):429–453.

Khalmetski, K. (2016). Testing guilt aversion with an exogenous shift in beliefs. *Games and Economic Behavior*, 97:110–119.

Khalmetski, K., Ockenfels, A., and Werner, P. (2015). Surprising gifts: Theory and laboratory evidence. *Journal of Economic Theory*, 159:163–208.

Knetsch, J. L. and Wong, W.-K. (2009). The endowment effect and the reference state: Evidence and manipulations. *Journal of Economic Behavior & Organization*, 71(2):407–413.

Kőszegi, B. and Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, 121(4):1133–1165.

Kőszegi, B. and Rabin, M. (2007). Reference-dependent risk attitudes. *American Economic Review*, 97(4):1047–1073.

Krupka, E. L., Leider, S., and Jiang, M. (2017). A meeting of the minds: informal agreements and social norms. *Management Science*, 63(6):1708–1729.

Le Quement, M. T., Patel, A., et al. (2018). Communication as gift-exchange. Technical report, School of Economics, University of East Anglia, Norwich, UK.

Loewenstein, G. and Molnar, A. (2018). The renaissance of belief-based utility in economics. *Nature Human Behaviour*, 2(3):166–167.

Loomes, G. and Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The economic journal*, 92(368):805–824.

Mannahan, R. (2019). Self-esteem and rational self-handicapping. *Unpub'lished*.

Miettinen, T. and Suetens, S. (2008). Communication and guilt in a prisoner's dilemma. *Journal of Conflict Resolution*, 52(6):945–960.

Morell, A. (2019). The short arm of guilt–an experiment on group identity and guilt aversion. *Journal of Economic Behavior & Organization*, 166:332–345.

Motro, D., Ordóñez, L. D., Pittarello, A., and Welsh, D. T. (2018). Investigating the effects of anger and guilt on unethical behavior: A dual-process approach. *Journal of Business Ethics*, 152(1):133–148.

Moulton, R. W., Burnstein, E., Liberty Jr, P. G., and Altucher, N. (1966). Patterning of parental affection and disciplinary dominance as a determinant of guilt and sex typing. *Journal of Personality and Social Psychology*, 4(4):356.

Nyborg, K. (2018). Social norms and the environment. *Annual Review of Resource Economics*.

Ockenfels, A. and Werner, P. (2014a). Beliefs and ingroup favoritism. *Journal of Economic Behavior & Organization*, 108:453–462.

Ockenfels, A. and Werner, P. (2014b). Scale manipulation in dictator games. *Journal of Economic Behavior & Organization*, 97:138–142.

O'Donoghue, T. and Sprenger, C. (2018). Reference-dependent preferences. In *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 1, pages 1–77. Elsevier.

Olthof, T. (2012). Anticipated feelings of guilt and shame as predictors of early adolescents' antisocial and prosocial interpersonal behaviour. *European Journal of Developmental Psychology*, 9(3):371–388.

Patel, A. and Smith, A. (2019). Guilt and participation. *Journal of Economic Behavior & Organization*, 167:279–295.

Peeters, R. and Vorsatz, M. (2021). Simple guilt and cooperation. *Journal of Economic Psychology*, 82:102347.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American economic review*, pages 1281–1302.

Regner, T. and Harth, N. S. (2014). Testing belief-dependent models. *Jena Economic Research Papers*.

Reinikka, R. and Svensson, J. (2004). Local capture: evidence from a central government transfer program in uganda. *The quarterly journal of economics*, 119(2):679–705.

Reuben, E., Sapienza, P., and Zingales, L. (2009). Is mistrust self-fulfilling? *Economics Letters*, 104(2):89–91.

Ross, L., Greene, D., and House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology*, 13(3):279–301.

Shoji, M. (2020). Guilt and antisocial conformism: Experimental evidence from bangladesh. Technical report, University Library of Munich, Germany.

Spiekermann, K. and Weiss, A. (2016). Objective and subjective compliance: A norm-based explanation of 'moral wiggle room'. *Games and Economic Behavior*, 96:170–183.

Steenhaut, S. and Van Kenhove, P. (2006). The mediating role of anticipated guilt in consumers' ethical decision-making. *Journal of business ethics*, 69(3):269–288.

Tangney, J. and Fisher, K. (1995). *Self-conscious emotions: the psychology of shame, guilt and pride*. New York: The Guilford Press.

Tangney, J. P., Dearing, R. L., Wagner, P. E., and Gramzow, R. (1989). Test of self-conscious affect–3.

Vanberg, C. (2008). Why do people keep their promises? an experimental test of two explanations 1. *Econometrica*, 76(6):1467–1480.

Wang, X. and McClung, S. R. (2012). The immorality of illegal downloading: The role of anticipated guilt and general emotions. *Computers in Human Behavior*, 28(1):153–159.

Contents

Remerciements **I**

Résumé de la thèse **i**

Motivation de la thèse i

Préférences dépendantes des croyances i

Aversion à la culpabilité v

Objectifs de la thèse ix

Etat de l'art xi

Avancées théoriques xi

Contributions méthodologiques xv

Aperçu de la thèse xxi

Aversion à la culpabilité et détournement de fonds xxi

Aversion à la culpabilité et vulnerabilité xxiii

Aversion à la culpabilité et acquisition d'informations xxv

Discussion de la thèse xxviii

Introduction **1**

Motivation 1

Belief-Dependent Preferences 1

Guilt Aversion 4

Objectives . . . 8
State of the Art . . . 10
Theoretical Advances . . . 10
Methodological Contributions . . . 14
Outline . . . 19
Guilt Aversion and Embezzlement . . . 19
Guilt Aversion and Vulnerability . . . 21
Guilt Aversion and Information Acquisition . . . 23

1 Guilt Aversion and Embezzlement **33**
1.1 Introduction . . . 33
1.2 Theoretical Model and Behavioral Hypotheses . . . 38
1.2.1 The Embezzlement Mini-Game(s) . . . 38
1.2.2 Utility Functions . . . 40
1.2.3 Theoretical Predictions . . . 43
1.2.3.1 Predictions on Donor's behavior . . . 44
1.2.3.2 Predictions on Intermediary's behavior . . . 47
1.3 Experimental Design and Procedures . . . 51
1.3.1 Experimental Design . . . 52
1.3.2 Procedures . . . 56
1.4 Results . . . 57
1.4.1 Donors' Behavior . . . 58
1.4.2 Intermediaries' Behavior . . . 60
1.4.3 A Structural Estimate of Guilt Sensitivity . . . 66
1.5 Discussion and Conclusion . . . 68

Appendices **76**
1.A Additional Results . . . 76

1.B Previous Literature . . . 81
1.C Instructions . . . 82
1.C.1 Instructions for the Lab Experiment [Translated from French] . . . 82
1.C.2 Instructions for the Online Questionnaire [Translated from French] 93

2 Guilt Aversion and Vulnerability **100**
2.1 Introduction . . . 100
2.2 The Quasi-Trust Mini-Games . . . 105
2.3 Theoretical Model and Hypotheses . . . 109
2.3.1 Utility Functions . . . 110
2.3.2 Best-Reply Analysis . . . 113
2.3.2.1 Player A's Best-Reply Functions . . . 113
2.3.2.2 Player B's Best-Reply Functions . . . 115
2.3.3 Hypotheses . . . 118
2.3.3.1 Hypotheses on Player A's Behavior . . . 118
2.3.3.2 Hypotheses on Player B's Behavior . . . 119
2.4 Experimental Design and Procedures . . . 122
2.4.1 Decisions and Elicitation of Beliefs . . . 122
2.4.2 Elicitation of Individual Preferences . . . 124
2.4.3 Procedures . . . 124
2.5 Results . . . 126
2.5.1 A-Subjects' Behavior . . . 126
2.5.2 B-Subjects' Behavior . . . 128
2.6 Conclusion . . . 137

Appendices **142**
2.A Literature . . . 142
2.B Player A's Best-Reply Functions and Hypotheses . . . 144

2.B.1 Player A's Best-Reply Functions . . . 144
2.B.2 Player A's Behavioral Hypotheses . . . 146
2.C Instructions (Translated from French) . . . 150
2.D Additional Results . . . 161
2.D.1 Summary Statistics on Participants, by Treatment . . . 161
2.D.2 Detailed Analysis of A-Subjects Behavior . . . 161
2.D.3 Within-Individual Analysis of B-Subjects' Patterns of Choices . 166

3 Guilt Aversion and Information Acquisition 171
3.1 Introduction . . . 171
3.2 Theoretical Model . . . 178
3.2.1 Belief Formation . . . 180
3.2.2 Belief-dependent preferences . . . 180
3.2.3 Information Acquisition Strategy . . . 183
3.3 Design . . . 186
3.4 Experimental Hypotheses . . . 190
3.5 Experimental Results . . . 192
3.5.1 Are beliefs affected by the outside option manipulation? . . . 192
3.5.2 Are trustees motivated by belief-dependent preferences? . . . 195
3.5.3 How do belief-based preferences affect information acquisition? 196
3.6 Discussion and Conclusion . . . 200

Appendices 207
3.A Proofs adapted from Spiekermann and Weiss (2016) . . . 207
3.A.1 Proof on the variation of $\hat{u}$ with respect to Φ . . . 207
3.A.2 Proof of Proposition 3 . . . 208
3.A.3 Proof of Proposition 4 . . . 210
3.B Screens from the online experiment . . . 212

3.B.1 Trustors' screens 212
3.B.2 Trustees' screens 219
3.C Additional Results 232
3.C.1 Trustor's behavior 232
3.C.2 Trustees' justification of their sampling strategies 232
3.C.3 Trustees' likelihood of having a given sampling strategy 233
3.C.4 Determinants of beliefs, returns and preference type. 234
3.D Robustness checks Restricted sample 235
3.D.1 Are beliefs affected by the outside option manipulation? 236
3.D.2 Are trustees motivated by belief-dependent preferences? 237
3.D.3 How do belief-based preferences affect information acquisition 238

Conclusion **239**

List of Tables

1.1 Correlation between the intermediaries' decisions and their Guilt-NBE and Altruism scores . . . 65
1.2 Structural estimates of the guilt-sensitivity parameter . . . 67
1.A.1 Summary statistics of participants per session . . . 76
1.A.2 Summary statistics on beliefs and social norms . . . 77
1.A.3 Donors' *Give* choices for a given FOB on the frequency of *Transfer* choices 78
1.A.4 Matching the donors' behavior to our predictions - (*Keep*, *Keep*) . . . 78
1.A.5 Matching the donors' behavior to our predictions - (*Give*, *Keep*) . . . 78
1.A.6 Matching the donors' behavior to our predictions - (*Give*, *Give*) . . . 78
1.A.7 Matching the donors' behavior to our predictions - (*Keep*, *Give*) . . . 79
1.A.8 Intermediaries' *Transfer* choices for a given induced SOB . . . 79
1.A.9 Regression on the decision to Transfer (Logit model, fixed effects) . . . 79
1.A.10 Regression on the decision to Transfer (Logit model, random effects) . 80
1.A.11 Correlation between recipients' beliefs and recipients' risk aversion . . 80
1.A.12 Difference in social norms distributions across roles . . . 80
1.B.1 Previous estimations of the proportion of guilt-averse individuals . . . 81
1.B.2 Previous estimations of the guilt-sensitivity parameter . . . 81
1.B.3 Previous correlation of personality traits and behavioral outcomes . . . 81
1.C.1 GASP Questionnaire - Answers Key . . . 94

1.C.2 Honesty-Humility Scale - Answers Key 95

2.3.1 A's predicted behavior depending on her altruism sensitivity ϕ_{Ah} and first-order belief α_{AB}, with altruistic (resp., selfish) strategy in dark grey (resp., light grey) color. 114

2.3.2 Hypotheses on the proportion of guilt-averse B-players in each game-treatment combination 121

2.5.1 Likelihood of B-Subjects Choosing *Right*, by Game 131

2.5.2 Structural Estimates of Guilt sensitivity for B-Subjects Disclosing Behavior Consistent with the Model 136

2.A.1 List of published experiments on guilt aversion 143

2.D.1 Summary Statistics on Participants, by Treatment 161

2.D.2 Proportion of *In* Choices Across Games, by First-Order Belief 162

2.D.3 Likelihood of A-Subjects Choosing *In*, by Games 163

3.5.1 Average marginal effects of monetary incentives on the likelihood of each sampling strategy 199

3.C.1 Trustees' justification of their sampling strategies 233

3.C.2 Average marginal effects of preferences types on the likelihood of each sampling strategy 234

3.C.3 Determinants of participants' beliefs, trustees' conditional return decisions and preferences type. 235

List of Figures

1.1 The Embezzlement Mini-Game(s) 39

1.2 Predicted behavior of a rational donor in the two conditions (Low, High), depending on his altruistic type (γ_D) and first-order belief (α_{DI}) 45

1.3 Predicted behavior of a rational intermediary for the four possible second-order beliefs $\beta_{Ij} \in \{0, 1/3, 2/3, 1\}$, depending on his guilt type (θ_{Ij}) and altruistic type (γ_I) 49

1.4 Distribution of the donors' choices depending on their first-order beliefs 58

1.5 Distribution of the intermediaries' switching second-order beliefs 61

1.6 Distribution of the intermediaries' choices depending on their induced second-order beliefs 62

1.C.1 Screenshot for the *Donor Treatment* 88

1.C.2 Screenshot for the *Recipient Treatment* 89

2.2.1 The Investment Game 105

2.2.2 The Reversed-Investment Game 105

2.2.3 The Donation Game 106

2.2.4 The Exploitation Game 106

2.2.5 Vulnerability in the four Quasi-Trust mini-games 109

2.5.1 Distribution of B-Subjects' Pattern of Choices Across Games and Treatments 129

2.C.1 Screenshot in Treatment A 156
2.C.2 Screenshot in Treatment C 156
2.D.1 Distribution of B-Subjects Consistency of Behavior 167

3.2.1 Trust game with High or Low outside option 179
3.3.1 Choosing a source of information 189
3.5.1 Distribution of trustors and trustees' beliefs about trustors' payoff from *In*. 193
3.5.2 Distribution of individual beliefs about trustors' expected payoff from *In* 194
3.5.3 Trustees' return strategies 195
3.5.4 Distribution of information acquisition strategies for belief-independent and guilt-averse trustees. 197
3.D.1 Distribution of trustors and trustees' beliefs about trustors' payoff from *In*. 236
3.D.2 Distribution of individual beliefs about trustors' expected payoff from *In* 236
3.D.3 Trustees' return strategies 237
3.D.4 Information acquisition strategy. 238

Introduction

"Guilt is a powerful sting"

Paul Auster, Leviathan (1992)

Motivation

Belief-Dependent Preferences

Bentham was the first to introduce the concept of utility function in 1789, a concept that is now at the core of the economic science. He considered that individuals seek to maximize pleasures and minimizes pains. Interestingly, pleasures and pains can correspond to beliefs (e.g, pleasure and pain from memory, or pleasure and pain from imagination). However, since this seminal contribution, economists have considered that individuals seek to maximize their expected payoff and have mostly considered beliefs as a constraint on individuals' expected payoff. It is only recently that experimental economists came back to a wider conception of what individuals seek to maximize; these economists challenged the assumption of an exclusively self-interested agent to incorporate other-regarding preferences in the utility function (for a review see Cooper, 2009). Yet, most of these new models (e.g. Fehr and Schmidt, 1999; Charness and Rabin, 2002) continue to define preferences based on outcomes (e.g., inequality aversion, efficiency concern). Moving in the direction of Bentham's suggestion, it is crucial to

also consider preferences based on beliefs. For instance, guilt aversion, the focus of this thesis, is a type of preference which leads individuals to avoid disappointing their beliefs about others' expectations. This belief-based preference can affect a variety of decisions. Indeed, the rise of environmental friendly policies may partly be explained by the guilt of policy makers who do not want to disappoint anymore the expectations of the young generations. Besides, the influence of guilt aversion contributes to account for the gap between some developing countries in which bribery is common and others where asking for a bribe is an exception. In the former, asking for a bribe is devoid of any guilt since it does not disappoint any expectations of integrity, and hence bribery persists. In the later, guilt aversion prevents bribery as there are high expectations of integrity. In our day-to-day life, deciding how much to tip a taxi-driver also depends on the guilt we may feel from disappointing the expectations of the taxi-driver.

Moving beyond guilt-averse preferences, belief-dependent preferences can help understand why policy-makers should not focus exclusively on the material wealth of a population to assess its happiness (Easterlin, 1995). Indeed, individuals' well-being may depend on how they (and others) see themselves, on their evaluation of their choices relative to their expectations, on how they value reciprocating others' intentions or on which emotions they feel. Each of these motives can be modelled by incorporating beliefs as preferences in the utility function.[15] Recent reviews by Battigalli and Dufwenberg (2020) and Loewenstein and Molnar (2018) illustrate how belief-dependent preferences are widespread. First, these preferences allow to account for the importance of image concerns. The image that individuals are trying to maintain relates either to their actions (I want others to believe that I did action X) or to their traits (I want myself/others to

[15]Tools to explore belief-dependent preferences have been developed in the psychological game theory framework. This framework was initially proposed by Geanakoplos et al. (1989) and further extended by Battigalli and Dufwenberg (2009). The key feature of this theoretical framework is to let utility at an end-node depends on beliefs, while in traditional game theory it cannot be the case. We observe a steady increase in the number of articles citing this theoretical framework from 1991 to 2017 (Azar, 2019).

believe that I am Y). For instance, an individual may dislike that an audience believes he/she cheated (Dufwenberg and Dufwenberg, 2018) or may dislike being perceived as self-interested (e.g., Bodner and Prelec, 2003; Bénabou and Tirole, 2006 ;Grossman and Van Der Weele, 2017). In general, this literature challenges the view that pro-social actions are made out of pure pro-social preferences, but shows that they also respond to the decision-makers' beliefs about their image. Another illustration is given by Mannahan (2019) model, cited by Battigalli and Dufwenberg (2020), which assumes that an individual's utility increases with his/her beliefs in his/her good traits (e.g., intelligence). This may lead an individual, who is going to be externally evaluated, to self-handicap before this evaluation in order to make the signal on his/her ability worthless (by fear of discovering his/her true ability).

Second, belief-dependent utility functions allow for expectation-based reference-dependent preferences, that is preferences that depend on the initial expectation hold by a decision-maker (for a survey see O'Donoghue and Sprenger, 2018). One of the most prominent examples of such preferences is given by the models of Kőszegi and Rabin (2006) and Kőszegi and Rabin (2007).[16] They proposed that individuals are averse to "disappointment", that is, to obtaining a payoff lower than expected. This framework offers an account for a variety of phenomena such as daily income targeting (Farber, 2005), consumer choices (Heidhues and Kőszegi, 2008), endowment effect (Knetsch and Wong, 2009), real effort in competition (Gill and Prowse, 2012) or job search (DellaVigna et al., 2017).

Third, the belief-dependent framework allows researchers to model intention-based reciprocal preferences (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004, 2019). These models contend that individuals like to reciprocate (un)kindness with (un)kindness, and that the kindness of an action is evaluated with respect to intentions. Applied works on

[16]Seminal works in the same vein were carried out by Loomes and Sugden (1982) and Bell (1985).

reciprocal preferences show that such preferences can explain phenomena as diverse as framing effects (Dufwenberg et al., 2011), trade disputes (Conconi et al., 2017), communication (Le Quement et al., 2018), or climate change negotiations (Nyborg, 2018).

Finally, belief-dependent preferences allow researchers to introduce emotions in utility functions. As noted by Elster (1998), economists have largely neglected the study of emotions. He argues that "*emotions are triggered by beliefs*" (p.49), which makes belief-based utility functions their natural home. The impact of emotions is two-fold. On the one hand, emotions can lead to a behavioral response (i.e., an action). For instance, based on the frustration-aggression hypothesis of Dollard et al. (1939), Battigalli et al. (2019) proposed a model of anger and frustration. They consider that individuals who are frustrated by their payoff (compared to their expectations) react with anger and may be willing to destroy the payoff of others. On the other hand, emotions can also affect behavior given that individuals anticipate the emergence of emotions (positive or negative) in planning their behavior. For instance, anticipating the anxiety of their patients can affect doctors' decision to reveal or not an unfavorable medical diagnosis (Caplin and Leahy, 2004). As mentioned earlier, the anticipation of disappointment, a negative emotion, can also influence the present behavior (Loomes and Sugden, 1982; Bell, 1985). We now consider more extensively the emotion which is the focus of this thesis, namely, guilt.

Guilt Aversion

In psychology, Baumeister et al. (1994) distinguished two functions of guilt. First, guilt being aversive, its *anticipation* will discourage harmful behavior. Second, when the harm is done, the *experience* of guilt encourages to repair the transgressor's relationship with

the other party and to avoid repeating the same harmful behavior in the future.

In economics, Battigalli and Dufwenberg (2007) introduced the concept of guilt aversion which corresponds to an aversive anticipation of guilt.[17] A decision-maker is guilt-averse toward another player if he/she incurs a disutility from disappointing the other player's expectations. This is in line with Baumeister et al. (1994) proposal stating that "*If people feel guilt for hurting their partners [...] and for failing to live up to their expectations, they will alter their behavior (to avoid guilt) in ways that seem likely to maintain and strengthen the relationship.*" It is important to note that Battigalli and Dufwenberg (2007) consider empirical expectations about how one *will* behave and not normative expectations about how one *ought* to behave.[18]

Guilt frequently occurs in the course of daily life. Results from Baumeister et al. (1995) revealed that, during a week, participants experienced guilt around 13% of their waking hours. In economics, the existence of guilt-aversion has been assessed in a variety of

[17] Psychologists have largely focused their research on the experience of guilt. Yet, notable exceptions exist to show that the anticipation of guilt can alter future behaviors. For example, when making the decision of whether or not to donate bone marrow, contemplating future feelings of guilt related to inaction can lead to donating (Massi Lindsey, 2005); anticipated guilt is the strongest predictor of organs donation's intention after controlling for norms, self-efficacy and family discussion (Wang, 2011). In another domain, Wang and McClung (2012) showed that anticipated guilt significantly influences college students' intentions to illegal download among those who already did so in the previous 6 months (but not among those who did not). Furthermore, adolescents who reported high anticipated guilt to aggressive behavior were, according to their peers, more likely to behave pro-socially and less likely to behave anti-socially (Olthof, 2012). Finally, Steenhaut and Van Kenhove (2006) found that enhancing the anticipation of guilt encouraged consumers' intentions to buy ethical products (on ethical decision making, see also Motro et al., 2018).

[18] This distinction relates to the difference between the guilt-aversion model and models of social norms such as d'Adda et al. (2016). However, Hauge (2016) and Krupka et al. (2017) showed that both injuctive and empirical expectations matter for explaining behavior in, respectively, a dictator game and informal agreements. Within the realm of empirical expectations, Danilov et al. (2019) distinguished between the action commonly chosen by other decision-makers in the same situations (social norm) and the action expected by another party directly affected by the choice of the decision-maker (guilt aversion). When information on both benchmarks were revealed simultaneously, the authors found that both types of information affected transfers in the dictator game

games. Guilt aversion has been shown to promote cooperation in trust games (e.g., Charness and Dufwenberg, 2006; Reuben et al., 2009; Bellemare et al. (2017); see Cartwright, 2019b for a review) and in sender-receiver games (e.g., Battigalli et al., 2013). Evidence of guilt-averse behavior was found notably in a credence good game (Beck et al., 2013), in a lost wallet game (Dufwenberg and Gneezy, 2000) and in a public good game (Patel and Smith, 2019). Guilt aversion also motivates prosocial behavior in dictator games (e.g., Ockenfels and Werner, 2014a; Hauge, 2016; Bellemare et al., 2018): the more the recipient expects to receive, the more the dictator gives.

Moreover, guilt aversion has been proposed to account for decisions in contexts more diverse than those corresponding to the usual economic games presented above. In a pioneer work, Dufwenberg (2002) studied the following situation: a wife first decides to support or not her husband (for instance while he studies), then the husband (who now have a degree) decides to divorce (reaping all the earnings) or to stay married (sharing the earnings). Dufwenberg (2002) showed that, if the husband's guilt parameter is sufficiently strong, the signal of trust by the wife induces enough guilt to "force" the husband to stay married. Balafoutas (2011) investigated a game between a bureaucrat who can be bribed, a lobby who aims at bribing the bureaucrat and the public who may suffer from the bribery, where the bureaucrat can feel guilty from disappointing the public's expectations. In a one-shot interaction, allowing for the bureaucrat to be guilt-averse reduces corruption. However, in repeated interactions, and under some conditions (e.g., when beliefs are updated relatively fast), the society may be entrapped in self-fulling beliefs of corruption. Further, examining the consequences of guilt aversion in a context of tax evasion, Dufwenberg and Nordblom (2018) assumed that a taxpayer may suffer from guilt from evading taxes. They showed that, when endogenizing the behavior of the tax authority, introducing guilt-averse preferences reduces the inspection rate (which involves savings of public funds). Finally, Cartwright (2019a) highlighted the importance

of beliefs in the decision to use drugs in sports. He found that under sufficiently high guilt, athletes have an incentive to race clean if others expect them to do so, irrespective of the behaviors of the other athletes in the race. It is noteworthy that these works are applied theoretical models that remain to be empirically tested.

Two other well-established phenomena, promise-keeping and in-group favouritism, have been traditionally accounted for by guilt aversion. Expectation-based account of promise-keeping propose that making a promise raises the expectations of the recipient, and the decision maker does not want to let down the recipient's expectations (Charness and Dufwenberg, 2006; Ederer and Stremitzer, 2017). Similarly, in-group favoritism could be explained, at least in part, by guilt aversion (rather than only by intrinsic in-group preferences). The guilt induced by in-group members' expectations differs from that of out-group members', either because these expectations are believed to be higher by the decision-maker (Güth et al., 2009; Guala et al., 2013; Ockenfels and Werner, 2014b) or because they are more valued by the decision maker (Morell, 2019).[19]

Guilt is prevalent in many types of situations and has been suggested as an explanation for a variety of phenomena. Yet, the question of the situational factors that modulate its influence remains under-investigated (Ghidoni and Ploner, 2020).[20] Following Balafoutas and Fornwagner (2017), the question is no more whether guilt aversion exists, but under which circumstances it is relevant.

[19] However, some experimental studies have suggested that guilt aversion could not account for promise-keeping (Vanberg, 2008; Di Bartolomeo et al., 2018) or in-group favoritism (Ciccarone et al., 2020).

[20] Two exceptions are notable. First, there is converging evidence that "reasonable" expectations are more likely to be taken into account by guilt-averse players (Regner and Harth, 2014; Khalmetski, 2016; Balafoutas and Fornwagner, 2017; Danilov et al., 2019). Second, Khalmetski (2016) and Bellemare et al. (2018) showed that dictators' average guilt sensitivity decreased in the level of stakes.

Objectives

The aim of this thesis is to question the scope of influence of guilt-aversion by investigating (i) the direction of guilt: can individuals be guilt averse toward someone who is not affected monetarily by their decisions, or only toward someone whose earnings are affected? (ii) some necessary conditions for the emergence of guilt aversion: does the vulnerability of the person to whom individuals may feel guilty impacts the emergence of guilt? (iii) the robustness of guilt aversion against self-serving bias: Are individuals strategic in their information acquisition about others' expectations (in order to avoid triggering their guilt)?

On the one hand, in order to show that the range of influence of guilt is broader than initially suspected, we challenged the assumption of Battigalli and Dufwenberg (2007) model which contends that one feels guilty only towards co-players who are monetarily affected. Experimentally, we designed three-players games that were never investigated before (Chapters 1 and 2). On the other hand, in order to test the "objective" vs. "illusory" nature of guilt aversion, we let participants be uninformed about others' expectations and we formalized how guilt-averse individuals should acquire this information (Chapter 3).

In Chapter 1, we considered a situation in which a donor could send money to a recipient. As in the case of charitable donations or government transfers to developing countries, this money had to be transmitted through an intermediary who could embezzle some of it. The behavior of this intermediary was the focus of our research. Whereas previous studies have exclusively considered potential guilt of the intermediary toward a recipient (as in dictator games), we believed that an intermediary could also experience guilt toward the donor who forewent part of his endowment to increase the recipient's payoff. Our research objective was two-fold. First, we aimed at documenting the existence of

guilt aversion among intermediaries both toward the recipient and the donor. Second, we wanted to test whether the direction of guilt, i.e., toward the donor or toward the recipient mattered. In fact, we extended Battigalli and Dufwenberg (2007) model to allow the intermediary to feel guilty toward the donor, who is not affected monetarily by the intermediary's action.

Aside from this topical case of the donor in embezzlement situations, in Chapter 2 we wondered, more generally, whether the vulnerability of the co-players was a necessary condition for the emergence of guilt. Investigating a new modulator of guilt seemed like a promising avenue to better understand its nature. Indeed, guilt aversion has been shown to be sensitive to modulators such as pre-play communication between players (*e.g.* Balafoutas and Sutter, 2017) or the reasonability of expectations (*e.g.* Balafoutas and Fornwagner, 2017). With regard to vulnerability, as previous studies were based on the model of Battigalli and Dufwenberg (2007), which primitive is that the dictator/trustee can feel guilty toward a vulnerable recipient/trustor, they were not able to address this question. To fill this gap, we designed four Quasi-Trust mini-games that systematically vary the vulnerability of the co-players.

Chapters 1 and 2 tested whether the scope of influence of guilt could be extended to new situations (e.g., when a co-player is not affected monetarily, or not vulnerable). In Chapter 3, we investigated whether the impact of guilt aversion on behavior might be overstated by previous experimental designs. Indeed, guilt aversion is considered as a pro-social preference, but this type of preference tends to be challenged by situational excuses (e.g., Dana et al., 2007). In fact, previous experimental paradigms, where the uncertainty about others' expectations is ultimately resolved when actions are implemented, leave little room for such excuses. In this chapter, we chose to leave the decision-maker

uncertain about others' expectations. We adapted the model of information acquisition by Spiekermann and Weiss (2016) in order to make predictions on how belief-dependent individuals acquire information on others' expectations. Then, an experiment allowed us to test these predictions and to detect whether some individuals self-servingly bias their information acquisition strategy in order to minimize the tension between their monetary interest and their belief-dependent motivation.

State of the Art

Theoretical Advances

In this section, we describe in more detail the model of Battigalli and Dufwenberg (2007). We then expose the different extensions of this seminal work that have been proposed in the literature. In their guilt-aversion model, Battigalli and Dufwenberg (2007) distinguished two concepts: simple guilt and guilt-from-blame. With the former, a player cares about how much he/she is letting down a co-player's expectations, whereas with the later, a player cares about how much a co-player can blame him/her for letting down this co-player's expectations.

We focus on the most common concept: simple guilt.[21] In Battigalli and Dufwenberg (2007), the decision maker i experiences the utility from his/her material payoff π_i, and the disutility from feeling guilty G_{ij} (Equation 10). Battigalli and Dufwenberg (2007) introduced the disappointment function D_j (Equation 11) which represents the difference, if positive, between the co-player's j initial expectation about his/her material payoff (α_j) and the amount given by the decision-maker. The decision maker's guilt is determined

[21] Papers on guilt-from-blame are rare, maybe because guilt-from-blame requires to reason with third- and fourth-order beliefs, which is cognitively demanding (for exceptions, see Charness and Dufwenberg, 2011; Beck et al., 2013).

by the share of guilt that can be attributed to his/her choice s_i: $D_j(s_i, s_j) - min_{s_{\tilde{i}}} D_j(s_i, s_{\tilde{i}})$. Finally, θ_i represents the guilt sensitivity parameter that is unique to each individual.

$$u_i(z, s_i, \alpha_j) = \pi_i(z) - G_{ij}(z, s_i, \alpha_j) \quad (9)$$

$$\text{where } G_{ij}(z, s_i, \alpha_j) = D_j(s_i, s_j) - min_{s_{\tilde{i}}} D_j(s_i, s_{\tilde{i}}) \quad (10)$$

$$\text{and } D_j = max\{E_{s_j, \alpha_j}[\pi_j] - \pi_j, 0\} \quad (11)$$

The seminal model of Battigalli and Dufwenberg (2007) has opened the way to different lines of extensions. First, Khalmetski et al. (2015) extended the simple guilt model to account for the joy of positive surprises and not only the dis-utility of negative surprises. Second, Inderst et al. (2019) nuanced the definition of guilt by allowing the possibility of blaming the co-player.

Khalmetski et al. (2015) developed an innovative extension of the Battigalli and Dufwenberg (2007) model to include the pleasure of bearing surprising gifts, and not only the guilt from bringing surprising setbacks. They proposed that decision makers may also like to surprise positively their co-players' expectations.[22] As explained in Equation 13, they considered the surprise function S_i where the first term represents the utility from positive surprises (when $x > t_i$) and the second term represent the disutility from negative surprises (when $x < t_i$). α_i and β_i correspond, respectively to the propensity to make positive surprises and avoid negative surprises. Finally, the reference point x is the distribution of first-order beliefs, given by the probability density function h_j.[23]

[22] Note that, they also extended their model to integrate the notion that the dictators may care for the recipients' inferences about the dictators' intentions (as in guilt from blame).

[23] Unlike many applied models of guilt aversion (e.g., Beck et al., 2013), Khalmetski et al. (2015) did not take the point expectation as the co-player's reference point. Rather, they consider that the co-player's

$$u_i(\pi_j, h_j) = \pi_i + S_i(\pi_j, h_j) \quad (12)$$

$$\text{where } S_i(\pi_j, h_j) = \alpha_i \int_0^{p_j} (\pi_j - x) h_j(x) dx - \beta_i \int_0^{p_j} (x - \pi_j) h_j(x) dx \quad (13)$$

In the dictator game, this extension predicts that dictators who have a strong preference for positive surprises will have their transfer decisions negatively correlated with the recipients' expectation, unlike simple guilt which predicts a positive correlation. Indeed, when the recipient's expectations are low, it leaves more room for the dictator to create a positive surprise. Their extension was also applied in the context of a public good game by Dhami et al. (2019). They found that 30% of belief-dependent dictators liked to positively surprise their recipients.

Although they proposed to widen the scope of influence of guilt aversion, Khalmetski et al. (2015) still considered surprises relative to the expectations about the co-player's own payoff. In the first two chapters in this thesis, we turn back to considering only negative surprises. Yet, based on three-player games, we extend the definition of guilt to the disappointment of expectations about another player's payoff, namely the most disadvantaged player's payoff.

To test the relevance of allowing to blame the disappointed co-player, Inderst et al. (2019) examined a customer-advisor game where the customer can decide between buying some verified information (*Out*) or trusting the advisor's advice (*In*). In this context, the authors introduced a new concept, coined as shared guilt, which captures the idea that "*the attribution of guilt for disappointing trust is shared between players whose choices*

reference point is stochastic: it corresponds to the probability distribution of the possible outcomes for the co-player.

eventually caused this disappointment, including the disappointed player herself". In a similar manner as Battigalli and Dufwenberg (2007) did for G_{ij}, Inderst et al. (2019) computed τ_i and τ_j as the share of disappointment that can be attributed to the decision maker's and, respectively, the co-player's j choices; they refer to τ_j as the self-blame of the co-player. The final guilt of the decision maker corresponds to a function increasing in his/her responsibility (in line with simple guilt) and decreasing in the co-player's self-blame (novelty of their shared guilt formulation).

$$U_i(\tau_i, \tau_j) = \pi_i - \theta_i G_i(\tau_i, \tau_j) \tag{14}$$

$$\text{where } \tau_i(s_i, s_j) = D_j(s_i, s_j) - min_{s_i} D_j(s_{\tilde{i}}, s_j) \tag{15}$$

$$\text{and } \tau_j(s_i, s_j) = D_j(s_i, s_j) - min_{s_{\tilde{j}}} D_j(s_i, s_{\tilde{j}}) \tag{16}$$

On the one hand, the model of simple guilt predicts that the higher the cost of information, the higher the customer's first-order beliefs conditional on *In*, hence, the lower the lying rate of the advisor. On the other hand, under shared guilt, the higher the cost of information, the higher the responsibility of the customer when choosing In, the higher the lying rate of the advisor. Inderst et al. (2019) found results in line with both predictions given the heterogeneity of the participants' preferences.

The study by Inderst et al. (2019) suggests that, when there is a way out of their guilty feelings, players take it. The third chapter of this thesis follows up on this intuition by allowing players to resolve the uncertainty about others' expectations in a strategic and self-serving manner.

Methodological Contributions

We will now describe and discuss the current methodologies used to capture guilt aversion. Before diving into the different approaches to capture guilt, it is important to note that guilt aversion has to be measured based on its anticipation rather than its actual expression. Indeed, as emphasized by Miettinen and Suetens (2008), guilt is a counterfactual emotion, that is an emotion induced by the thought of a defection that has not been realized yet. If we were to measure the feeling of guilt after an actual defection, rather than the anticipation of guilt, we would capture only the guilt of participants with the lowest guilt sensitivity, since the participants with the highest guilt sensitivity would have avoided incurring the psychological cost of defecting.

We first discuss two different approaches, with and without transmitting beliefs, that aim at evidencing the existence of guilt averse behaviors. However, guilt aversion is not an all-or-none emotion. Hence, we then consider studies that have attempted to estimate the degree of guilt.

There exist two main approaches to capture guilt averse behaviors: with and without transmitting beliefs. When transmitting beliefs, we distinguish between three methods: the *baseline* method (asking second-order beliefs), the *disclosure* method (disclosing first-order beliefs) and the *menu* method (conditioning the choice on possible first-order beliefs). In their pioneer study, Charness and Dufwenberg (2006) used the *baseline* method. They asked the participants about their second-order beliefs, that is, what they expect their co-players expect from them. Consistent with the guilt-aversion model, their results evidenced a positive correlation between the participants' second-order beliefs and their choices of how much they give to their co-players. However, the observed correlation may have been caused by false consensus effect (Ross et al., 1977): decision

makers tend to believe that others will think and act the same way as they do. Hence, decision makers may have predicted their co-players' expectations based on what they intended to do. To control for this potential false-consensus effect, Ellingsen et al. (2010) proposed an experimental design in which participants were asked their first-order beliefs about the decision maker's behavior. Then, without having said so in advance, these first-order beliefs were transmitted to the co-player, to exogenously manipulate his/her second-order beliefs. This corresponds to the *disclosure* method. The authors did not find a positive correlation between the transmitted first-order beliefs and the players' behavior and they concluded that previous evidence did not capture guilt aversion but rather the extent of the false consensus effect. Khalmetski et al. (2015) reconciled those findings by making the distinction between correlations at the aggregate level and correlations at the individual level. As noted by Tangney and Fisher (1995) and allowed by Battigalli and Dufwenberg (2007) model, the degree to which individuals are sensitive to guilt may vary across individuals. Therefore, a null correlation observed at the aggregate level may reflect the fact that some individuals have a positive correlation while others have a negative correlation. To capture guilt-aversion at the individual level, they introduced the *menu* method which consisted in asking decision makers to formulate a series of choices conditional on possible first-order beliefs of their co-player. The choice actually implemented for payment is the one corresponding to the actual first-order belief of the co-player. This method has since been widely used by other experimentalists (e.g., Attanasi et al., 2013; Hauge, 2016; Balafoutas and Fornwagner, 2017; Bellemare et al. (2017); Bellemare et al. (2018); Dhami et al. (2019)). Comparing the different studies, we observe that both the *baseline* method and the *menu* method allowed the authors to evidence the presence of guilt-averse behaviors. Interestingly, Bellemare et al. (2017) reached the same conclusion in a within-study comparison of the two methods.[24] Finally, both the *baseline* and the *menu* method withhold some information from the participants.

[24] In fact, Bellemare et al. (2017) systematically compared the three methods previously cited. They also found that the *disclosure* method induces a higher unconditional level of kindness.

As underlined by Khalmetski et al. (2015) themselves,"*dictators might get suspicious when learning, before making their choices, that recipients were not informed about all strategically relevant aspects of the decision situation. This might create the impression that there are also possibly other aspects of the design that are withheld from the dictators.*" To answer this criticism, Khalmetski et al. (2015) designed a robustness check experiment in which, after eliciting the recipients' first-order beliefs, the authors told the participants that dictators would condition their choices on the recipients' first-order beliefs. Previous results were replicated, thereby suggesting that the dictators' choices were not influenced by suspicion.

Another approach to capture guilt-averse behaviors consists in exogenously manipulating the first-order beliefs of the co-player and providing full information on this manipulation to the decision maker. This results in manipulating the decision-maker second-order beliefs in the same direction as the beliefs of the co-player. Then, the experimentalist can observe whether the variation in beliefs leads to a variation in behavior consistent with guilt aversion without requiring transmitting actual beliefs. Within this approach, various experiments using different designs have supported the guilt aversion hypothesis.[25] The first instance of such methodology was proposed by Ederer and Stremitzer (2017). In a trust game, they introduced a random device which determined whether the trustee could decide how much to give to the trustor or whether the computer decided that the trustor received zero. The trustor knew ex-ante whether this random device was reliable (high likelihood to let the trustee decide) or unreliable (low likelihood to let the trustee decide). Building on the same idea, Khalmetski (2016) designed a sender-receiver game where the sender's material incentives to lie were either low or high. The receiver knew the

[25] Yet, as mentioned in the previous section, Ederer and Stremitzer (2017) found evidence of guilt aversion only in settings where it existed a direct promising link between the trustee and the trustor. Similarly, Balafoutas and Sutter (2017) found that their proxy for beliefs predicted transfers of the current dictator only when pre-play communication occurred.

ex-ante probability that the sender's incentives to lie were high. Finally, in a trust game, Inderst et al. (2019) manipulated the outside option of the trustor: the higher the outside option, the more the trustor was willing to forego by choosing *In*, the higher the trustor's first-order beliefs. In a similar vein, Balafoutas and Sutter (2017) revealed the recipients' history of past transfers to the current dictator. These past transfers served as a proxy to account for the recipients' expectations.

The first two chapters of this thesis employed the *menu* method where players can condition their choices on the possible expectations of the co-players, which allowed us to capture guilt aversion at the individual level. In Chapter 3, we implemented a combination of the *menu* method and the manipulation of the outside option in a trust game: players could condition their choices on the possible outside options of the co-players, and hence on their possible expectations. The interest of such methodology lies in the fact that the co-players' expectations are motivated by their outside option and have to be inferred by the decision-maker. These two reasons probably converge in reducing the probability that the decision-maker down-play these expectations.

Beyond evidencing the existence of guilt averse behaviors, researchers have attempted to estimate the guilt sensitivity parameter (θ_i in the simple guilt model).[26] Again, various methodologies have been used: (i) structural estimation, (ii) inferences from equilibrium predictions, (iii) inferences from information bounds or (iv) hypothetical questionnaires.

[26] While going beyond the dichotomic view of evidencing (or not) guilt aversion is a promising approach, it has also been developed by psychologists using guilt proneness questionnaires: the Guilt and Shame Proneness questionnaire developed by Cohen et al. (2011), the Test of Self-Conscious Affect developed by Tangney et al. (1989) or a single question proposed by Moulton et al. (1966). However, studies testing the consistency between the guilt sensitivity measured by economists and the guilt proneness measured through questionnaire are unconclusive. Bellemare et al. (2019) found a positive and strong correlation between the guilt proneness assessed by the Test of Self-Conscious Affect and the estimated guilt sensitivity parameter whereas Peeters and Vorsatz (2021) reported an absence of correlation between the guilt parameter and the guilt proneness assessed by the Guilt and Shame Proneness questionnaire.

Bellemare et al. (2011) used a structural model to estimate the Willingness-to-Pay to avoid disappointing the expectations of co-player in a dictator game.[27] Bellemare et al. (2018) updated their own estimation by allowing guilt sensitivity to depend on stakes size, which accounted for 60% of dictators' behaviors in their experiment. Second, Peeters and Vorsatz (2021) computed the set of all equilibria in a prisoner dilemma (i.e. cooperation rate) as a function of the guilt parameter. Then, they estimated the guilt parameter observed in the population from the experimental cooperation rates. In different participation games, Patel and Smith (2019) used a similar technique to compute the guilt sensitivity parameter based on the theoretical predictions of mixed symmetric equilibria. Third, Bellemare et al. (2019) proposed to infer information bounds on the guilt parameter without data or assumptions about beliefs.[28] Finally, Attanasi et al. (2016) and Peeters and Vorsatz (2021) proposed hypothetical methods to elicit the guilt sensitivity parameter. In a trust mini-game, Attanasi et al. (2016) asked the second-movers to consider a situation in which the first-mover has chosen *In* and they have chosen to send back nothing. Then, they elicited how much the second-movers are willing to pay back to the first-mover, for each possible first-order beliefs of the first-mover.[29] In a prisoner dilemma, Peeters and Vorsatz (2021) asked participants which was the minimum amount for which they would be indifferent between cooperating and defecting.

In Chapters 1 and 2, we followed Bellemare et al. (2011) in using a structural model to estimate the guilt sensitivity of the decision-makers. This was critical in our approach which aimed at diversifying contexts in which we can observe guilt aversion. Indeed, by

[27] Bellemare et al. (2011) experimental design involved one treatment where participants were asked to state their second-order beliefs and one treatment where participants were informed of their co-player first-order beliefs. The estimated guilt sensitivity parameter was significantly greater in the former treatment than the later, suggesting the presence of a false consensus effect.

[28] Unfortunately, their analysis yielded implausibly high estimates of guilt aversion, which were most likely due the very small proportion of players having this preference in their experiment.

[29] On a similar note, Chang et al. (2011) elicited the amount of guilt participants would have felt if they had returned less money in a trust game

doing so, we were able, not only to assess the existence of guilt in different contexts, but also to evidence whether its intensity varies across situations.

Outline

The literature reviewed up to now has shown that guilt aversion does exist. Yet, we do not know much about the factors that enhance or diminish its prevalence or impact. Investigating some of these factors is the focus of this thesis. In Chapter 1, within the topical situation of embezzlement, we investigated whether an individual can feel guilty toward a player who is not affected monetarily. In Chapter 2, we extended this question to multiple scenarios and varied in a systematic manner the vulnerability of the other players. In Chapter 3, we investigated whether the impact of guilt aversion on behavior may be "illusory" by studying individuals' behavior when they have the possibility to strategically acquire information about others' beliefs.

Guilt Aversion and Embezzlement

In Chapter 1, we considered a situation where a donor can send money to a recipient. As in the case of charitable donations or government transfers to developing countries, this money had to be transmitted through an intermediary who could embezzle some of it. Whereas previous research has exclusively considered the guilt of the intermediary toward a recipient (as in dictator games), we believed that an intermediary could also experience guilt toward the donor who donated this amount with the intention of giving to the recipient. We intended to test whether the direction of the guilt (toward the recipient or toward the donor) affects the intensity or the prevalence of guilt-aversion among intermediaries.

We designed a novel three-player game, the Embezzlement Mini-Game. In this game, a donor sends a donation to a recipient, but this donation has to be transferred by an intermediary who can embezzle a fraction of this donation to increase his own material payoff. The donor forms expectations on how much of the donation the intermediary will transfer to the recipient; and the recipient forms expectations on how much he/she will receive. Hence, the intermediary may disappoint two types of expectations. To capture the intermediary's aversion to disappoint expectations, i.e., his/her guilt, we allowed the intermediary to condition his/her transfer decision on the expectations of either the donor (Donor treatment) or the recipient (Recipient treatment). This manipulation was made between-subjects and it allowed us to capture guilt-aversion at the individual level: intermediaries who increased their transfers with their co-player's expectations were classified as guilt-averse. Moreover, we varied within-subjects the percentage of the donation that could be embezzled (80% in the High condition and 60% in the Low condition) to test the extent to which the intensity of potential embezzlement affected beliefs.

Regarding the guilt toward the recipient, we relied on Battigalli and Dufwenberg (2007) definition of guilt as the disutility from letting down the recipient's expectations about his own payoff. Regarding the guilt toward the donor, we extended theoretically Battigalli and Dufwenberg (2007) model. Rather than not letting down the donor's expectations about his own payoff (which is not affected by the decision to embezzle), an intermediary guilt-averse toward the donor dislikes letting down the donor's expectations about another player's material payoff, i.e., the recipient.

As predicted, we found that guilt aversion reduces embezzlement among intermediaries. Furthermore, our experimental results indicated that the proportion of intermediaries who experience guilt is similar regardless of whether the guilt is toward the recipient

or toward the donor (on average, 25%). The absence of difference across treatments is confirmed when looking at the intensity of their aversions (on average, an intermediary is willing to pay 0.37 ECU to not let down another player by 1 ECU). This is striking as embezzlement affects the earnings of the recipient but not those of the donor. It shows that the mechanisms at play in guilt aversion extend in situations where decisions have no direct monetary consequences. These results hold irrespective of the percentage of the donation that could be embezzled.

Guilt Aversion and Vulnerability

In Chapter 1, we have demonstrated that an individual (the intermediary) can feel guilty even toward someone who is not affected monetarily (the donor). Building on this finding, we then explored the role of vulnerability in triggering one's guilt.

Based on the findings of the previous chapter, in Chapter 2, we distinguished two types of vulnerability. On the one hand, an individual may be vulnerable "ex-post" if his/her final payoff depends on the action of the decision-maker. On the other hand, an individual can be vulnerable "ex-ante" if his/her initial endowment can be entrusted to the decision-maker. In this chapter, we aimed at assessing whether guilt aversion is moderated by the different combinations of the two types of vulnerability of the player towards whom the decision-maker may feel guilty. Furthermore, we intended to test whether the two types of vulnerability are complements or substitutes in their impact on guilt.

To address these questions, we designed a laboratory experiment with four two-stage games with two active players (A and B) and one passive player (C). In each game, the second-mover (B) can be entrusted by the first-mover (A) with a sum of money. This money comes from the endowment of either player A or C, depending on the game– ex-

ante vulnerability of player A or C. Then, player B can redistribute this money between himself and another player (A or C, depending on the game) – ex-post vulnerability of player A or C. These four mini-games varied systematically the vulnerability of players A and C. They were played within-subjects in a random order. Furthermore, we compared between-subjects whether player B's behavior was elicited conditional on the first-order beliefs of player A or C. By doing so, we manipulated whether player B's guilt aversion is elicited toward a player whose intentions are observable (player A, active) or not (player C, passive).

Theoretically, we follow the model exposed in Chapter 1 which contends that guilt is activated even when the beliefs of the disappointed player do not concern her material payoff but the payoff of a third player. We went one step further by allowing player B's guilt aversion to be triggered even when player A's intention was mediated by another player's first-order beliefs, namely the passive player. In other words, our theoretical model predicts that guilt-aversion depends neither on the game, nor on the status of the disappointed player (active *vs* passive).

We found evidence of guilt averse behaviours in all four games. Incidentally, this revealed the relevance of guilt aversion in games where it was never tested before. In line with our theoretical predictions, our results showed no significant difference neither in the proportion of guilt-averse players B nor in their guilt intensity. Some players B even exhibited a guilt-averse behavior toward players that were not vulnerable in either dimension. Finally, observing or not the intention of the co-players does not seem to modulate the guilt aversion of the decision-maker.

Guilt Aversion and Information Acquisition

Chapter 2 has extended the scope of guilt aversion by revealing its existence in a variety of situations where it was never tested before. Chapter 3 addressed another limitation of previous studies and challenged the strength of this preference by using a design where players can self-select the information about others' expectations, which allowed us to explore strategies used by individuals to avoid feeling guilty while behaving selfishly.

A large body of evidence showed that individuals care about the welfare of others (for a survey see Cooper, 2009). Yet, these apparently pro-social preferences have been shown to fade away in the presence of uncertainty about the relationship between one's actions and outcomes (e.g., Dana et al., 2007). In contrast, little is known about the robustness of these preferences when uncertainty concerns others' expectations. In this chapter, we tackled this question by investigating whether individuals with belief-dependent preferences self-servingly biased their information acquisition strategy in order to minimize the tension between their monetary interest and their belief-dependent motivation.

We adapted the information acquisition model by Spiekermann and Weiss (2016) to study whether guilt-averse and reciprocal agents strategically acquire information about others' expectations. This approach allows to distinguish between objective and subjective preferences. Applied to the domain of belief-dependent preferences, objective preferences imply that agents maximize their utility by complying to others' actual expectations whereas subjective preferences imply that agents maximize their utility by complying to the belief they have about others' expectations (subjective preferences). Our model predicted that agents with objective belief-dependent preferences always prefer more information regardless of their belief-dependent motivation, while agents with subjective belief-dependent preferences strategically seek information that mini-

mizes the trade-off between their monetary interest and their belief-dependent motivation.

We have tested our predictions in an online experiment. We designed a modified trust game in which we manipulated the trustor's outside option in order to influence the trustors' first-order beliefs and, by forward induction, the trustees' second-order beliefs. In fact, we created an exogeneous variation in beliefs for 61.25% of trustees in our sample. We then elicited trustees' preferences by asking them to report their return choices conditionally on learning the trustor's outside option. We found that 52.04% of trustees were belief-independent, 43.88% of subjects were guilt-averse and 4.08% were reciprocal. Finally, trustees were given the unexpected opportunity to acquire information about the trustor's outside option. Importantly, the implemented choice of the trustees depended on the information they had: certainty that the outside option is Low, certainty that the outside option is High or uncertainty about the outside option.

We found that the majority of belief-dependent trustees exhibited an information acquisition strategy consistent with subjective preferences: 60.47% of guilt-averse trustees sought a Low signal only. Both regression analyses and answers to the post-experimental questionnaire suggest that this choice was made to maximize one's own payoff. Symmetrically, the only reciprocal trustee in our sample sought a High signal only. Finally, it is worth mentioning that a non-trivial fraction of our sample acquired information in a pattern consistent with objective belief-dependent preferences (20.93%). Our results suggest that the positive impact of belief-dependent preferences on pro-social choices depends on the (un)certainty about other players' beliefs.

To conclude, the whole ambition of this thesis was to challenge the scope of influence of guilt-aversion. Our work revealed for the very first time that guilt aversion can also

be triggered when the person to whom one feels guilty is not affected monetarily. This finding, obtained in the context of an embezzlement game, was further extended to new contexts where guilt aversion was systematically observed toward players independently of their vulnerability. Finally, although guilt aversion seemed to generalize to a variety of situations, we demonstrated that its robustness may be challenged in situations where the decision-makers have the possibility to avoid the tension between their monetary incentives and their belief-dependent concerns.

Bibliography

Attanasi, G., Battigalli, P., and Manzoni, E. (2016). Incomplete-information models of guilt aversion in the trust game. *Management Science*, 62(3):648–667.

Attanasi, G., Battigalli, P., and Nagel, R. (2013). Disclosure of belief-dependent preferences in a trust game. Technical report, No. 506, IGIER (Innocenzo Gasparini Institute for Economic Research), Bocconi University.

Azar, O. H. (2019). The influence of psychological game theory. *Journal of Economic Behavior & Organization*, 167:445–453.

Balafoutas, L. (2011). Public beliefs and corruption in a repeated psychological game. *Journal of Economic Behavior & Organization*, 78(1-2):51–59.

Balafoutas, L. and Fornwagner, H. (2017). The limits of guilt. *Journal of the Economic Science Association*, 3(2):137–148.

Balafoutas, L. and Sutter, M. (2017). On the nature of guilt aversion: Insights from a new methodology in the dictator game. *Journal of Behavioral and Experimental Finance*, 13:9–15.

Battigalli, P., Charness, G., and Dufwenberg, M. (2013). Deception: The role of guilt. *Journal of Economic Behavior & Organization*, 93:227–232.

Battigalli, P. and Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, 97(2):170–176.

Battigalli, P. and Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory*, 144(1):1–35.

Battigalli, P. and Dufwenberg, M. (2020). Belief-dependent motivations and psychological game theory.

Battigalli, P., Dufwenberg, M., and Smith, A. (2019). Frustration, aggression, and anger in leader-follower games. *Games and Economic Behavior*, 117:15–39.

Baumeister, R. F., Reis, H. T., and Delespaul, P. A. (1995). Subjective and experiential correlates of guilt in daily life. *Personality and Social Psychology Bulletin*, 21(12):1256–1268.

Baumeister, R. F., Stillwell, A. M., and Heatherton, T. F. (1994). Guilt: an interpersonal approach. *Psychological bulletin*, 115(2):243.

Beck, A., Kerschbamer, R., Qiu, J., and Sutter, M. (2013). Shaping beliefs in experimental markets for expert services: Guilt aversion and the impact of promises and money-burning options. *Games and Economic Behavior*, 81:145–164.

Bell, D. E. (1985). Reply—putting a premium on regret. *Management Science*, 31(1):117–122.

Bellemare, C., Sebald, A., and Strobel, M. (2011). Measuring the willingness to pay to avoid guilt: estimation using equilibrium and stated belief models. *Journal of Applied Econometrics*, 26(3):437–453.

Bellemare, C., Sebald, A., and Suetens, S. (2017). A note on testing guilt aversion. *Games and Economic Behavior*, 102:233–239.

Bellemare, C., Sebald, A., and Suetens, S. (2018). Heterogeneous guilt sensitivities and incentive effects. *Experimental Economics*, 21(2):316–336.

Bellemare, C., Sebald, A., and Suetens, S. (2019). Guilt aversion in economics and psychology. *Journal of Economic Psychology*, 73:52–59.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American economic review*, 96(5):1652–1678.

Bodner, R. and Prelec, D. (2003). Self-signaling and diagnostic utility in everyday decision making. *The psychology of economic decisions*, 1(105):26.

Caplin, A. and Leahy, J. (2004). The social discount rate. *Journal of political Economy*, 112(6):1257–1268.

Cartwright, E. (2019a). Guilt aversion and reciprocity in the performance-enhancing drug game. *Journal of Sports Economics*, 20(4):535–555.

Cartwright, E. (2019b). A survey of belief-based guilt aversion in trust and dictator games. *Journal of Economic Behavior & Organization*, 167:430–444.

Chang, L. J., Smith, A., Dufwenberg, M., and Sanfey, A. G. (2011). Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron*, 70(3):560–572.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6):1579–1601.

Charness, G. and Dufwenberg, M. (2011). Participation. *American Economic Review*, 101(4):1211–37.

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

Ciccarone, G., Di Bartolomeo, G., and Papa, S. (2020). The rationale of in-group favoritism: An experimental test of three explanations. *Games and Economic Behavior*, 124:554–568.

Cohen, T. R., Wolf, S. T., Panter, A. T., and Insko, C. A. (2011). Introducing the gasp scale: a new measure of guilt and shame proneness. *Journal of personality and social psychology*, 100(5):947.

Conconi, P., DeRemer, D. R., Kirchsteiger, G., Trimarchi, L., and Zanardi, M. (2017). Suspiciously timed trade disputes. *Journal of International Economics*, 105:57–76.

Cooper, D. (2009). Other regarding preferences: a survey of experimental results in j. kagel & a. roth. *The handbook of experimental economics*, 2.

d'Adda, G., Drouvelis, M., and Nosenzo, D. (2016). Norm elicitation in within-subject designs: Testing for order effects. *Journal of Behavioral and Experimental Economics*, 62:1–7.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.

Danilov, A., Khalmetski, K., and Sliwka, D. (2019). Descriptive norms and guilt aversion. *Mimeo*.

DellaVigna, S., Lindner, A., Reizer, B., and Schmieder, J. F. (2017). Reference-dependent job search: Evidence from hungary. *The Quarterly Journal of Economics*, 132(4):1969–2018.

Dhami, S., Wei, M., and al Nowaihi, A. (2019). Public goods games and psychological utility: Theory and evidence. *Journal of Economic Behavior & Organization*, 167:361–390.

Di Bartolomeo, G., Dufwenberg, M., Papa, S., and Passarelli, F. (2018). Promises, expectations & causation. *Games and Economic Behavior*.

Dollard, J., Miller, N. E., Doob, L. W., Mowrer, O. H., and Sears, R. R. (1939). Frustration and aggression.

Dufwenberg, M. (2002). Marital investments, time consistency and emotions. *Journal of Economic Behavior & Organization*, 48(1):57–69.

Dufwenberg, M. and Dufwenberg, M. A. (2018). Lies in disguise–a theoretical analysis of cheating. *Journal of Economic Theory*, 175:248–264.

Dufwenberg, M., Gächter, S., and Hennig-Schmidt, H. (2011). The framing of games and the psychology of play. *Games and Economic Behavior*, 73(2):459–478.

Dufwenberg, M. and Gneezy, U. (2000). Measuring beliefs in an experimental lost wallet game. *Games and economic Behavior*, 30(2):163–182.

Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and economic behavior*, 47(2):268–298.

Dufwenberg, M. and Kirchsteiger, G. (2019). Modelling kindness. *Journal of Economic Behavior & Organization*, 167:228–234.

Dufwenberg, M. and Nordblom, K. (2018). Tax evasion with a conscience.

Easterlin, R. A. (1995). Will raising the incomes of all increase the happiness of all? *Journal of Economic Behavior & Organization*, 27(1):35–47.

Ederer, F. and Stremitzer, A. (2017). Promises and expectations. *Games and Economic Behavior*, 106:161–178.

Ellingsen, T., Johannesson, M., Tjøtta, S., and Torsvik, G. (2010). Testing guilt aversion. *Games and Economic Behavior*, 68(1):95–107.

Elster, J. (1998). Emotions and economic theory. *Journal of economic literature*, 36(1):47–74.

Farber, H. S. (2005). Is tomorrow another day? the labor supply of new york city cabdrivers. *Journal of political Economy*, 113(1):46–82.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3):817–868.

Geanakoplos, J., Pearce, D., and Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and economic Behavior*, 1(1):60–79.

Ghidoni, R. and Ploner, M. (2020). When do the expectations of others matter? experimental evidence on distributional justice and guilt aversion. *Theory and Decision*, pages 1–46.

Gill, D. and Prowse, V. (2012). A structural analysis of disappointment aversion in a real effort competition. *American Economic Review*, 102(1):469–503.

Grossman, Z. and Van Der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association*, 15(1):173–217.

Guala, F., Mittone, L., and Ploner, M. (2013). Group membership, team preferences, and expectations. *Journal of Economic Behavior & Organization*, 86:183–190.

Güth, W., Ploner, M., and Regner, T. (2009). Determinants of in-group bias: Is group affiliation mediated by guilt-aversion? *Journal of Economic Psychology*, 30(5):814–827.

Hauge, K. E. (2016). Generosity and guilt: The role of beliefs and moral standards of others. *Journal of Economic Psychology*, 54:35–43.

Heidhues, P. and Kőszegi, B. (2008). Competition and price variation when consumers are loss averse. *American Economic Review*, 98(4):1245–68.

Inderst, R., Khalmetski, K., and Ockenfels, A. (2019). Sharing guilt: How better access to information may backfire. *Management Science*, 65(7):3322–3336.

Khalmetski, K. (2016). Testing guilt aversion with an exogenous shift in beliefs. *Games and Economic Behavior*, 97:110–119.

Khalmetski, K., Ockenfels, A., and Werner, P. (2015). Surprising gifts: Theory and laboratory evidence. *Journal of Economic Theory*, 159:163–208.

Knetsch, J. L. and Wong, W.-K. (2009). The endowment effect and the reference state: Evidence and manipulations. *Journal of Economic Behavior & Organization*, 71(2):407–413.

Kőszegi, B. and Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, 121(4):1133–1165.

Kőszegi, B. and Rabin, M. (2007). Reference-dependent risk attitudes. *American Economic Review*, 97(4):1047–1073.

Krupka, E. L., Leider, S., and Jiang, M. (2017). A meeting of the minds: informal agreements and social norms. *Management Science*, 63(6):1708–1729.

Le Quement, M. T., Patel, A., et al. (2018). Communication as gift-exchange. Technical report, School of Economics, University of East Anglia, Norwich, UK.

Loewenstein, G. and Molnar, A. (2018). The renaissance of belief-based utility in economics. *Nature Human Behaviour*, 2(3):166–167.

Loomes, G. and Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The economic journal*, 92(368):805–824.

Mannahan, R. (2019). Self-esteem and rational self-handicapping. *Unpub'lished.*

Massi Lindsey, L. L. (2005). Anticipated guilt as behavioral motivation: An examination of appeals to help unknown others through bone marrow donation. *Human Communication Research*, 31(4):453–481.

Miettinen, T. and Suetens, S. (2008). Communication and guilt in a prisoner's dilemma. *Journal of Conflict Resolution*, 52(6):945–960.

Morell, A. (2019). The short arm of guilt–an experiment on group identity and guilt aversion. *Journal of Economic Behavior & Organization*, 166:332–345.

Motro, D., Ordóñez, L. D., Pittarello, A., and Welsh, D. T. (2018). Investigating the effects of anger and guilt on unethical behavior: A dual-process approach. *Journal of Business Ethics*, 152(1):133–148.

Moulton, R. W., Burnstein, E., Liberty Jr, P. G., and Altucher, N. (1966). Patterning of parental affection and disciplinary dominance as a determinant of guilt and sex typing. *Journal of Personality and Social Psychology*, 4(4):356.

Nyborg, K. (2018). Social norms and the environment. *Annual Review of Resource Economics*.

Ockenfels, A. and Werner, P. (2014a). Beliefs and ingroup favoritism. *Journal of Economic Behavior & Organization*, 108:453–462.

Ockenfels, A. and Werner, P. (2014b). Scale manipulation in dictator games. *Journal of Economic Behavior & Organization*, 97:138–142.

O'Donoghue, T. and Sprenger, C. (2018). Reference-dependent preferences. In *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 1, pages 1–77. Elsevier.

Olthof, T. (2012). Anticipated feelings of guilt and shame as predictors of early adolescents' antisocial and prosocial interpersonal behaviour. *European Journal of Developmental Psychology*, 9(3):371–388.

Patel, A. and Smith, A. (2019). Guilt and participation. *Journal of Economic Behavior & Organization*, 167:279–295.

Peeters, R. and Vorsatz, M. (2021). Simple guilt and cooperation. *Journal of Economic Psychology*, 82:102347.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American economic review*, pages 1281–1302.

Regner, T. and Harth, N. S. (2014). Testing belief-dependent models. *Jena Economic Research Papers*.

Reuben, E., Sapienza, P., and Zingales, L. (2009). Is mistrust self-fulfilling? *Economics Letters*, 104(2):89–91.

Ross, L., Greene, D., and House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology*, 13(3):279–301.

Spiekermann, K. and Weiss, A. (2016). Objective and subjective compliance: A norm-based explanation of 'moral wiggle room'. *Games and Economic Behavior*, 96:170–183.

Steenhaut, S. and Van Kenhove, P. (2006). The mediating role of anticipated guilt in consumers' ethical decision-making. *Journal of business ethics*, 69(3):269–288.

Tangney, J. and Fisher, K. (1995). *Self-conscious emotions: the psychology of shame, guilt and pride*. New York: The Guilford Press.

Tangney, J. P., Dearing, R. L., Wagner, P. E., and Gramzow, R. (1989). Test of self-conscious affect–3.

Vanberg, C. (2008). Why do people keep their promises? an experimental test of two explanations 1. *Econometrica*, 76(6):1467–1480.

Wang, X. (2011). The role of anticipated guilt in intentions to register as organ donors and to discuss organ donation with family. *Health communication*, 26(8):683–690.

Wang, X. and McClung, S. R. (2012). The immorality of illegal downloading: The role of anticipated guilt and general emotions. *Computers in Human Behavior*, 28(1):153–159.

Chapter 1

Guilt Aversion and Embezzlement

This chapter is co-authored with Giuseppe Attanasi and Marie Claire Villeval. It was published in the *Journal of Economic Behavior & Organization* in 2019.

1.1 Introduction

Contrary to the standard economics-of-crime approach that focuses on the trade-off between the monetary costs and benefits of dishonesty (Becker and Stigler, 1974; Fan et al., 2009), the behavioral economic analysis of unethical behavior insists on the importance of incorporating moral costs in this trade-off. Indeed, not all individuals cheat, even when there is no risk of detection, and most cheaters do not exploit their opportunity of cheating maximally, which may come from the moral costs associated with unethical behavior (see, *e.g.*, Abeler et al., 2019; Fischbacher and Föllmi-Heusi, 2013; Gneezy, 2005; Kajackaite and Gneezy, 2017; Mazar et al., 2008, in the context of lying, and Abbink and Serra, 2012; Drugov et al., 2014; Köbis et al., 2016, in the context of corruption). However, little is known on the nature of these moral costs beyond the idea that most people are willing to maintain a positive self-image. Psychological game

theory, introduced by Geanakoplos et al. (1989) and further developed by Battigalli and Dufwenberg (2009), helps to understand the nature of these moral costs through the modeling of emotions such as guilt aversion, although this theory has been rarely used so far to investigate dishonesty (for a recent exception, see Dufwenberg and Dufwenberg, 2018).

In this chapter, we study guilt aversion in the context of embezzlement. Embezzlement is defined as the misappropriation of assets by individuals to whom they were entrusted, in order to monopolize or to steal them. It can occur when the providers of resources need intermediaries to transfer these resources to the final recipients. The problem is crucial, especially in developing countries, in domains such as health, education, or humanitarian aid where the final recipients seldom receive the totality of aid transfers they are entitled to.[1] Indeed, donors must rely on local intermediaries and usually cannot verify which amount has eventually been transferred to the entitled recipients. Embezzlement is detrimental to economic development and cooperation (Beekman et al., 2014; Olken and Pande, 2012) and it can result in some programs becoming inequality enhancing (*e.g.*, Reinikka and Svensson, 2004) or no longer cost-effective (*e.g.*, Ferraz et al., 2012).

While most of the previous literature has studied interventions affecting the monetary costs and benefits attached to embezzlement (*e.g.*, Barr et al., 2009; Di Tella and Schargrodsky, 2003; Olken, 2007), we investigate the moral cost of embezzling by studying the intermediary's willingness to avoid the anticipated negative valence associated with guilt from embezzlement. Guilt aversion implies that an agent suffers a cost, *i.e.*, feels guilty, if he lets down others' expectations (Tangney and Fisher, 1995). Our first research

[1] For example, in 2000 in Ghana, a Public Expenditure Tracking Survey revealed that 80% of non-salary funds did not reach health facilities (Canagarajah and Ye, 2001). For the period 1991-1995, Ugandan schools received on average 13% of the governmental transfers they were entitled to (Reinikka and Svensson, 2004). In 2013, the head of the governmental High Relief Committee was arrested for the misappropriation of US$ 10 million earmarked for the aid of refugees in Lebanon.

objective is to identify in the laboratory the existence of such guilt aversion and its impact on the behavior of intermediaries who can embezzle the donations made by donors to recipients. Our second objective is to test whether the *direction of guilt aversion* matters, *i.e.*, whether it is stronger toward the donor or toward the recipient.

We designed a novel three-player game – the Embezzlement Mini-Game. In this game, a donor sends a donation to a recipient but this donation has to be transferred by an intermediary who can embezzle a fraction of this donation to increase his own material payoff.[2] Embezzlement decreases both the utility of the donor who cares about the recipient's well-being and the utility of the recipient who receives the donation. The donor forms expectations on how much of the donation the intermediary will transfer to the recipient. The recipient also has expectations on how much he will receive. Depending on his decision, the intermediary can fulfill or not the other two players' expectations. Hence, the intermediary may be affected by *donor-guilt aversion* and by *recipient-guilt aversion*.

Indirect evidence of intermediaries' guilt aversion can be found in previous studies. Chlaß et al. (2015) found that the more intermediaries believe that donors have donated, the more they transfer. This is coherent with our model's intuition which predicts that the more donors believe the donation will be transferred, the more intermediaries transfer. Di Falco et al. (2016) found that intermediaries at the beginning of longer transfer chains embezzle less than intermediaries in short chains. Feeling guilty from letting down the

[2]The game is meant to represent a situation in which an individual in a rich country sends money to a charity to help improve the situation of individuals in need in developing countries. The donor and the charity have to rely on local intermediaries, *e.g.*, the heads of villages. In many cases, these intermediaries are in a position to embezzle part of the donations – see for example the field experiments reported in Beekman et al. (2014).

recipients' expectation could explain this behavior.

We rely on the modeling of simple guilt aversion as a belief-dependent motivation by Charness and Dufwenberg (2006) and Battigalli and Dufwenberg (2007) in the framework of psychological game theory. This theory departs from traditional game theory in assuming that players' utilities do not only depend on their decisions but also on their beliefs about decisions, beliefs, or information. In particular, the psychological utility of a guilt-averse player depends on his second-order beliefs, *i.e.*, his beliefs about other players' beliefs about his own decision. For the recipient-guilt aversion of the intermediary, we rely on Battigalli and Dufwenberg (2007) definition of guilt as the disutility from letting down the recipient's expectations about *his own* payoff. For the donor-guilt aversion of the intermediary, we extend theoretically Battigalli and Dufwenberg (2007) model of simple guilt by introducing a novelty in the definition of guilt in the psy-games literature. Rather than not letting down the donor's expectations about his own payoff – which is not affected by the decision to embezzle –, a donor-guilt averse intermediary dislikes letting down the donor's expectations about *another player's* material payoff, *i.e.*, the recipient's payoff. In this case, the psychological utility of the guilt-averse player (the intermediary) depends on his beliefs about another player's beliefs (the donor) on a third player's material payoff (the recipient).[3]

Our theoretical analysis builds on the incomplete-information framework with role-dependent guilt of Attanasi et al. (2016). We assume that among the two active players only the intermediary can feel guilty. We enrich the set of psychological types by assuming that both the donor and the intermediary have altruistic preferences toward the

[3] In Balafoutas (2011) proposed a model that allows for an official who accepts a bribe to feel guilty toward both the citizen and the lobby. However, both directions of guilt are coherent with Battigalli and Dufwenberg (2007) model since the official can affect the payoff of both the citizen and the lobby.

recipient. Unlike Attanasi et al. (2016), we elaborate our behavioral hypotheses relying on best-reply analysis rather than on Bayesian equilibrium. This is motivated by the fact that a standard equilibrium analysis has no compelling foundation for games played one-shot (like ours) and in experiments on other-regarding preferences. Furthermore, and more importantly, in a psychological type space with the donor's and intermediary's altruism toward the recipient and with the intermediary's guilt toward the recipient and the donor, best-reply analysis can be carried out regardless of (in)completeness of information about players' types.[4] Thus, it delivers sharp predictions on the correlation between the intermediary's guilt types and behavior, and between his second-order beliefs and behavior, independently of the direction of guilt aversion (donor-guilt or recipient-guilt). Predicting the sign and size of these correlations is enough to provide appropriate behavioral hypotheses given the two research objectives mentioned above.

We implemented our Embezzlement Mini-Game in a laboratory experiment that allows us to measure directly the role of second-order beliefs on the intermediaries' decision to embezzle donations, adapting the belief-dependent menu method of Khalmetski et al. (2015). Between-subjects, we manipulated the information given to the intermediaries before they made their decision. In the Donor treatment, intermediaries decided whether to transfer or not the whole donation for each possible first-order belief of the donor on their decision. In the Recipient treatment, they made a decision for each possible first-order belief of the recipient on their decision. We can therefore compare the intermediaries' donor-guilt aversion and recipient-guilt aversion. Within-subjects, we manipulated the percentage of the donation that could be embezzled (80% in the High condition and 60% in the Low condition) to test how the intensity of potential embezzlement affects beliefs.

[4] Indeed, ours is an incomplete-information framework with private values. Hence, beliefs about the types of others do not enter the best-reply correspondence.

Our results show that on average 25% of the intermediaries are guilt-averse, *i.e.*, their decision to embezzle is influenced by others' expectations, and this holds regardless of the direction of the guilt and of the percentage of the donation that could be embezzled. Structural estimates indicate no difference in the effect of guilt aversion toward the donor and toward the recipient on intermediaries' behavior. This shows that guilt aversion may influence behavior even when decisions have no direct monetary consequences on the person toward whom guilt is directed.

The remainder of this paper is organized as follows. Section 1.2 introduces the theoretical model and its predictions. Section 1.3 describes the experimental design. Section 1.4 presents the results. Section 1.5 discusses and concludes.

1.2 Theoretical Model and Behavioral Hypotheses

1.2.1 The Embezzlement Mini-Game(s)

The Embezzlement Mini-Game involves three players: a donor, an intermediary and a recipient (see Figure 1.1). Players' material payoffs in Figure 1.1 are shown according to such order.

The three players receive an initial endowment: 150 ECU (Experimental Currency Units) for the donor, 80 ECU for the intermediary, and 10 ECU for the recipient (with 10 ECU = €1.2 in the experiment). Thus, the intermediary's endowment is the median between the donor's and the recipient's endowments.[5]

[5]The intermediary can be seen as the middleman in a network linking a NGO or a Governmental Agency to villagers. Different sets of possible actions for the donor and the intermediary capture asymmetry of positions. Unlike in a consecutive three-person dictator game (Bahr and Requate, 2014), the different initial endowments underline the different status of each player.

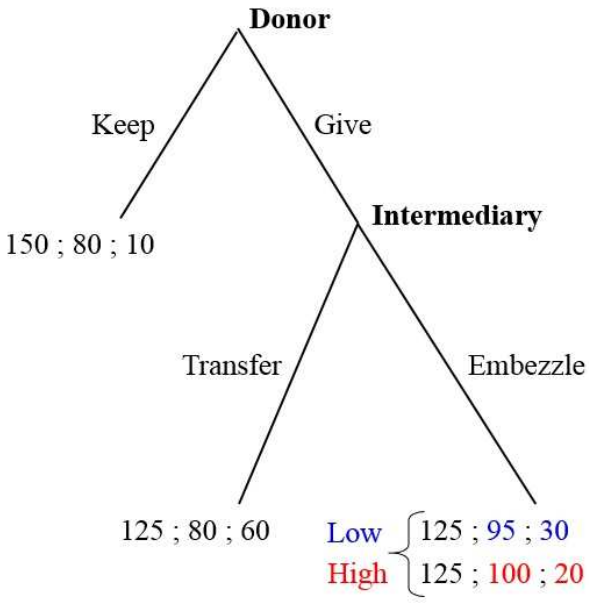


Figure 1.1: The Embezzlement Mini-Game(s)

The donor can *Keep* his endowment (in which case the game ends and each player earns his endowment) or *Give* 25 ECU to the recipient. However, the donation cannot be given *directly* to the recipient, it has to be transferred through the intermediary. The intermediary has to decide whether to *Transfer* the entirety of the donation to the recipient or to *Embezzle* a fraction f of the 25 ECU and transfer (1–f) to the recipient. The recipient receives twice the amount actually transferred. Thus, embezzlement involves an efficiency loss.[6]

Using a within-subject design, the Mini-Game is played under two conditions, each one allowing the intermediary to *Embezzle* a different fraction of the donation: in the Low condition, $f = 0.6$, and in the High condition, $f = 0.8$. Therefore, the two Mini-Games only differ for the set of possible actions of the intermediary (respectively, $f \in \{0, 0.6\}$

[6]This feature (also used in Boly et al. (2016)) captures a negative externality associated with embezzlement (see Ferraz et al. (2012), for an illustration in the domain of education in Brazil). The presence of a negative externality should reinforce the immoral image of embezzlement.

and $f \in \{0, 0.8\}$).

Figure 1.1 also shows two features of the final payoff distributions under each of these two conditions. First, no decision can lead to the equalization of payoffs between two or three players. Hence, no payoff distribution should be more salient than others. Second, the ranking of payoffs cannot be affected by the players' decisions. By doing so, we limit social comparison motives.

1.2.2 Utility Functions

Figure 1.1 shows respectively the *D*onor's, *I*ntermediary's and *R*ecipient's material payoff (M_j, with $j \in \{D, I, R\}$) at each terminal node of the Embezzlement Mini-Game, *i.e.*, for each profile of donor's and intermediary's strategy, respectively $s_D \in \{Keep, Give\}$ and $s_I \in \{Transfer \text{ if } Give, Embezzle \text{ if } Give\}$. We denote the donor's strategies *Keep* and *Give* with respectively K and G, and the intermediary's strategies $Transfer$ if *Give* and $Embezzle$ if *Give* with respectively T and E.

First of all, we assume that the **recipient's utility function** coincides with his material payoff, *i.e.*, $U_R(s_D, s_I) = M_R(s_D, s_I)$, which is made of his initial endowment, and the amount received $r(s_D, s_I)$. The latter enters the recipient's utility function only if the donor chooses *Give*, *i.e.*, $s_D = G$. In that case, the amount received depends on the amount actually transferred by the intermediary, $r(s_I)$. We are interested in the recipient's beliefs only in terms of their psychological impact on the intermediary's strategy s_I. Thus, in the experiment we only elicited α_{RI}, namely the recipient's first-order belief that the intermediary chooses *Transfer*, conditional on the donor choosing *Give*. Hence, our focus is on the recipient's expected received amount in this subgame:

$$\mathbb{E}_R[r(s_I)|s_D = G] = \alpha_{RI} \cdot r(T) + (1 - \alpha_{RI}) \cdot r(E) \tag{1.1}$$

Let us now introduce the **donor's utility function**. It is composed of his material payoff and his feeling of altruism toward the recipient (Eq. (1.2)). We assume that the donor (as well as the intermediary) have altruistic preferences toward the recipient. A player's feeling of altruism, A_{jR} with $j \in \{D,I\}$, represents player j's utility derived from an increase in the amount received by the recipient (*belief-independent* preferences). It is the product of two terms: $\gamma_j \geq 0$, player j's altruism sensitivity toward the recipient, *i.e.*, his altruistic type, and $r(s_D,s_I)$, the amount actually received by the recipient. With this, the donor's utility is:

$$U_D(\gamma_D,s_D,s_I) = M_D(s_D)+A_{DR}(\gamma_D,s_D,s_I) \tag{1.2}$$

$$\text{where } A_{DR}(\gamma_D,s_D,s_I) = \gamma_D \cdot r(s_D,s_I) \tag{1.3}$$

When the donor chooses between *Keep* and *Give*, he does not know what would be the intermediary's strategy. Therefore, his first-order belief that the intermediary chooses *Transfer* after *Give*, α_{DI}, matters for his giving choice. His expected utility conditional on choosing *Give* is:

$$\mathbb{E}_D[U_D(\gamma_D,s_I)|s_D = G] = M_D(G)+\gamma_D \cdot \mathbb{E}_D[r(s_I)|s_D = G] \tag{1.4}$$

where the amount the donor expects the recipient to get after his *Give* choice is:

$$\mathbb{E}_D[r(s_I)|s_D = G] = \alpha_{DI} \cdot r(T)+(1-\alpha_{DI}) \cdot r(E) \tag{1.5}$$

We made the simplifying assumption that a donor prefers that his donation increases the recipient's payoff rather than the intermediary's. This is broadly consistent with other models of distributional preferences: inequity aversion (Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999), since the recipient is the most disadvantaged player, and concern for efficiency (Charness and Rabin, 2002), since the sum of payoffs is maximized if the donor Gives and the intermediary Transfers. Importantly, our experimental

design allows us to test this assumption: we elicit the donor's first-order belief that the intermediary will choose *Transfer* after *Give*. If the donor's utility increases with the recipient's payoff, we should find that the frequency of giving increases in the donor's first-order belief about *Transfer* after *Give*.

We finally introduce the **intermediary's utility function**. The intermediary's utility (Eq. (1.6)) is composed of his material payoff M_I, his feeling of altruism toward the recipient A_{IR} (Eq. (1.7)), and his feeling of guilt toward the other players B_{Ij}, with $j \in \{D,R\}$ (Eq. (1.8)). As anticipated, we assume that the intermediary has altruistic preferences toward the recipient and that they are modeled as the donor's ones (see Eqs. (1.2) and (1.3)). Second, in line with the role-dependent guilt model of Attanasi *et al.* (2016), we assume that only the intermediary can feel guilty.[7] The intermediary's feeling of guilt, B_{Ij}, with $j \in \{D,R\}$, represents his disutility derived from letting down other players' expectations on the strategy he will select (*belief-dependent* preferences). It is the product of two terms: $\theta_{Ij} \geq 0$, the guilt sensitivity toward player $j \in \{D,R\}$, *i.e.*, the intermediary's guilt type; and the difference, if positive, between player j's expectations on the transferred amount after *Give*, $\mathbb{E}_j[r(s_I)|s_D = G]$, and the amount actually transferred to the recipient $r(s_I)$. This difference depends both on the intermediary's strategy, and on player j's first-order belief about this strategy (see Eqs. (1.1) and (1.5), respectively for $j = R$ and $j = D$).

If $\mathbb{E}_j[r(s_I)|s_D = G] > r(s_I)$, then the intermediary feels guilty from letting down player j's expectations on the amount transferred to the recipient.

[7] See the discussion in Attanasi et al. (2016), p. 649, where they argue that role dependence of guilt preferences is plausible in asymmetric games (see, *e.g.*, Attanasi et al., 2013, 2018, for indirect experimental evidence corroborating the assumption). In particular, they discuss how the assumption that sensitivity to guilt is triggered only when playing in the role of trustee (and not in the role of trustor) in the Trust Game resonates with the evolutionary psychology of emotions and the conceptual act theory of emotion. Similar arguments can be provided in support of sensitivity to guilt being triggered only when playing in the role of intermediary (and not in the role of donor) in the Embezzlement Mini-Game.

Independently from the treatment, the intermediary does not feel guilty if $(s_D, s_I) = (G, T)$, *i.e.*, the donor gives and the intermediary transfers the whole donation to the recipient. With this, the intermediary's utility is, for $j \in \{D, R\}$:

$$U_I(\theta_{Ij}, \gamma_I, s_I, \alpha_{jI} | s_D = G) = M_I(G, s_I) + A_{IR}(\gamma_I, s_I) - B_{Ij}(\theta_{Ij}, s_I, \alpha_{jI}) \quad (1.6)$$

$$\text{where } A_{IR}(\gamma_I, s_I) = \gamma_I \cdot r(s_I) \quad (1.7)$$

$$\text{and } B_{Ij}(\theta_{Ij}, s_I, \alpha_{jI}) = \theta_{Ij} \cdot \max\{0, \mathbb{E}_j[r(s_I)|s_D = G] - r(s_I)\} \quad (1.8)$$

Two clarifications are in order about Equation (1.8).

First, we analyze the impact of each guilt sensitivity (toward the donor and toward the recipient) separately because of our experimental design. We use a between-subject design to elicit the intermediary's belief-dependent strategy conditional on either the donor's (Donor treatment) or the recipient's (Recipient treatment) first-order beliefs. Therefore, we make the auxiliary assumption that one direction of guilt prevails over the other in each treatment, *i.e.*, $\theta_{IR} = 0$ in the Donor treatment and $\theta_{ID} = 0$ in the Recipient treatment.

Second, for B_{IR} (Eq. (1.8) with $j = R$), we rely on BD's (2007) definition of simple guilt as the intermediary's disutility from letting down the recipient's expectations about *his own* material payoff, whereas, for B_{ID} (Eq. (1.8) with $j = D$), we extend BD (2007) by defining the intermediary's guilt as the disutility from letting down the donor's expectations about *the recipient*'s material payoff ($\mathbb{E}_D[r(s_I)|s_D = G]$).

1.2.3 Theoretical Predictions

We provide a best-reply analysis of the Embezzlement Mini-Game(s) with incomplete information. We assume for simplicity that the recipient is commonly known to be selfish. But neither the donor's altruistic type, γ_D, nor the intermediary's guilt-altruistic type,

(θ_{Ij},γ_I), are known to the co-players.

The analysis relies on the assumption of players' rationality: each player is rational, *i.e.*, a subjective expected utility maximizer.[8]

1.2.3.1 Predictions on Donor's behavior

We define the donor's *Willingness-to-Give function* (WG) as the difference between his (expected) utility from *Give* (Eq. (1.5)) and his (certain) utility from *Keep*, the latter coinciding with his initial endowment, $M_D(K)$. The more the donor prefers to *Give* rather than *Keep*, the higher his willingness to *Give*:

$$\begin{aligned} WG(\gamma_D,\alpha_{DI}) &= \mathbb{E}_D[U_D(\gamma_D,s_I|s_D=G]-U_D(\gamma_D|s_D=K) \\ &= M_D(G)-M_D(K)+\gamma_D\cdot(\alpha_{DI}\cdot r(T)+(1-\alpha_{DI})\cdot r(E)) \end{aligned} \tag{1.9}$$

In the two conditions of the Mini-Embezzlement Game in Figure 1.1, a rational donor prefers to *Give* in the Low condition if $WG=-25+\gamma_D\cdot(\alpha_{DI}\cdot 30+20)>0$, and in the High condition if $WG=-25+\gamma_D\cdot(\alpha_{DI}\cdot 40+10)>0$. This leads to the following set of 'type-belief' pairs consistent with a *R*ational donor choosing *Give* in the Low and High conditions, respectively:

$$R_D^{G|Low}=\left\{(\gamma_D,\alpha_{DI}):\alpha_{DI}\geq\frac{1}{6}\left(\frac{5}{\gamma_D}-4\right)\right\} \tag{1.10}$$

$$R_D^{G|High}=\left\{(\gamma_D,\alpha_{DI}):\alpha_{DI}\geq\frac{1}{8}\left(\frac{5}{\gamma_D}-2\right)\right\} \tag{1.11}$$

Eqs. (1.10) and (1.11) are represented in Figure 1.2. The figure shows the $(s_D|Low,s_D|High)$ regions of the donor's 'type-belief' space (γ_D,α_{DI}), where the rational donor is predicted

[8] A two-step rationalizability procedure based on forward-induction reasoning (*cf.* Battigalli and Dufwenberg (2009), Section 5; Battigalli et al., 2019a,b) would provide similar qualitative predictions, by assuming that $\theta_{ID}>0$ and $\theta_{IR}>0$ at the same time in both treatments. Technical details of this analysis are available from the authors upon request.

to: *Keep* in both conditions (white region); *Give* in the Low and *Keep* in the High condition (light grey region); *Give* in both conditions (dark grey region). From Eqs. (1.10) and (1.11), it is easy to check that $(R_D^{K|Low} \cap R_D^{G|High}) = \varnothing$, *i.e.*, holding the 'type-belief' pair constant across conditions, a *R*ational donor cannot choose *Keep* in the Low and *Give* in the High condition, which explains the absence of a $(Keep, Give)$ region in Figure 1.2. The horizontal lines indicate the four possible first-order beliefs about the intermediary that a donor can hold in our experiment, $\alpha_{DI} \in \{0, 1/3, 2/3, 1\}$, as we will explain in Section 3.3.

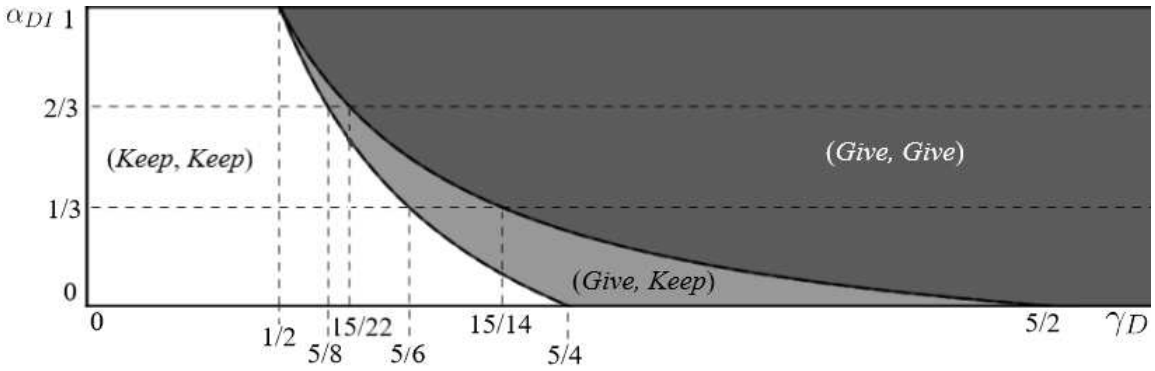


Figure 1.2: Predicted behavior of a rational donor in the two conditions (Low, High), depending on his altruistic type (γ_D) and first-order belief (α_{DI})

A comparative static analysis across the three regions of predictions in Figure 1.2 allows us to elaborate our hypotheses about the donor's behavior.

First, let us fix the pair $(\gamma_D, \alpha_{DI} = 0)$, *i.e.*, a donor with no trust on the intermediary's *Transfer* choice, and let us increase his first-order belief α_{DI}. For $\gamma_D \in [0, 1/2]$, the donor prefers to *Keep* in both conditions, for any α_{DI} (white region). For $\gamma_D \in (1/2, 5/4]$, as α_{DI} begins to increase, the donor switches from *Keep* to *Give* in the Low condition (light grey region); if α_{DI} continues to increase, he switches from *Keep* to *Give* also in the High condition (dark grey region). For $\gamma_D \in (5/4, 5/2]$, the donor prefers to *Give* in the Low condition, for any α_{DI} (light grey region); as α_{DI} increases, the donor switches

from *Keep* to *Give* in the High condition (dark grey region). For $\gamma_D \in (5/2, +\infty)$, the donor prefers to *Give* in both conditions, for any α_{DI} (dark grey region). Therefore, independently from the condition, an increase in α_{DI} never leads to a switch from *Give* to *Keep* and for some subset of donor's sensitivities to altruism it leads to a switch from *Keep* to *Give*. Considering heterogeneity in donors' types, we elaborate a hypothesis about the donor's **belief-dependent behavior**, whose verification is crucial to validate our assumption that the donor's utility increases with the recipient's payoff.

H.D1 [Choice-Belief Correlation]: The frequency of *Give* choices by altruistic donors increases in their first-order belief about *Transfer*.

Now suppose that the pair (γ_D, α_{DI}) is the same in the Low and High conditions, and refer again to Figure 1.2. If this 'type-belief' pair belongs to the white or the dark grey region, the rational choice is the same in both conditions, while if it lies in the light grey region, the rational choice is *Give* in the Low and *Keep* in the High condition. There is no 'type-belief' pair in the light grey region for $\alpha_{DI} = 1$, *i.e.*, when the donor is certain that the intermediary will *Transfer*. In that case, being the payoff profile after history (*Give*, *Transfer*) invariant to the condition (see 1.1), the donor's WG in Eq. (1.9) is the same both in the Low and High condition, and so the predicted choice. Since conditions are manipulated within-subjects (see Section 3.3), we assume that the distribution of donors' types is the same across conditions. Belief elicitation in the two conditions will allow us to check their invariance to the condition, that we assume in order to elaborate a hypothesis about the donor's **condition-dependent behavior**.

H.D2 [High *vs.* Low Condition on Choice]: Given the same donor's first-order belief about *Transfer* lower than one, same in both conditions, the frequency of *Give* choices by altruistic donors is higher in the Low than in the High condition.

Finally, let us fix the pair $(\gamma_D = 0, \alpha_{DI})$, *i.e.*, a selfish donor. Figure 1.2 shows that as his sensitivity to altruism γ_D increases, the donor moves from the white region directly to the dark grey region for $\alpha_{DI} = 1$, or passing through the light grey region for all $\alpha_{DI} < 1$. Therefore, independently from the condition, an increase in γ_D never leads to a switch from *Give* to *Keep* and it can lead to a switch from *Keep* to *Give*. Considering heterogeneity in donors' sensitivity to altruism, we elaborate a hypothesis about the donor's **type-dependent behavior**.

H.D3 [Choice-Type Correlation]: For a given first-order belief about *Transfer*, the frequency of *Give* choices increases with the donor's sensitivity to altruism.

Note that we derived H.D1, H.D2, and H.D3 without specifying the treatment (Donor or Recipient) since, in our experiment, donors are unaware of the treatment when they make their choices. Therefore, the donor's behavior should be **treatment-independent**.

1.2.3.2 Predictions on Intermediary's behavior

Relying on Eqs. (1.6–1.8), we define for each treatment (Donor and Recipient) the intermediary's *Willingness-to-Transfer function* (*WT*) as the difference between his utility when he *Transfers* and his utility when he *Embezzles*. Both these terms are expected utilities since the intermediary forms beliefs about the first-order beliefs α_{jI} of the co-player j toward whom he feels guilty ($j = D$ in the Donor and $j = R$ in the Recipient treatment).[9] These are his conditional second-order beliefs $\beta_{Ij} = \mathbb{E}_I[\alpha_{jI}|s_D = G]$, *i.e.*, for $j \in D, R$, conditional on the donor choosing *Give*.[10] The more the intermediary prefers

[9] Here we assume that the intermediary best-responds *as if* he had truly observed the donor's move. This holds by standard expected-utility maximization, except for the cases where the intermediary is certain that the donor has chosen *Keep*. Thus, we need the additional assumption that the intermediary has a belief conditional on *Give*, even if he is certain of *Keep*. Indeed, in our experiment the intermediary's decision is made under the strategy method, *i.e.*, both when the donor has chosen *Keep* and when he has chosen *Give* (see Section 3.3).

[10] More precisely, we reason as if the intermediary has a point belief β_{Ij} about α_{jI} conditional on *Give*.

to *Transfer* rather than *Embezzle*, the higher his willingness to *Transfer*.[11] Thus, for $j \in \{D,R\}$:

$$\begin{aligned} WT(\theta_{Ij},\gamma_I,\beta_{jI}|s_D=G) &= \mathbb{E}_I[U_I(\theta_{Ij},\gamma_I,\alpha_{jI}|G,T)]-\mathbb{E}_I[U_I(\theta_{Ij},\gamma_I,\alpha_{jI}|G,E)] \\ &= M_I(G,T)-M_I(G,E)+\gamma_I\cdot[r(T))-r(E)]+ \\ &\quad \theta_{Ij}\cdot[\beta_{Ij}\cdot r(T)+(1-\beta_{Ij})\cdot r(E))-r(E)] \end{aligned} \tag{1.12}$$

Rationality of the intermediary implies, for $j \in \{D,R\}$, that type $(\theta_{Ij},\gamma_I))$ with belief β_{Ij} chooses to *Transfer* the donation if $WT > 0$ in Eq. (2.6) and to *Embezzle* a fraction of it otherwise. In the two conditions of the Mini-Embezzlement Game in Figure 1.1, Eq. (2.6) becomes, for $j \in \{D,R\}$, $WT = -15+30\cdot(\gamma_I+\theta_{Ij}\cdot\beta_{Ij})$ in the Low condition and $WT = -20+40\cdot(\gamma_I+\theta_{Ij}\cdot\beta_{Ij})$ in the High condition. This leads to the following set of 'type-belief' pairs consistent with a *R*ational intermediary choosing *Transfer* in the Low and High conditions:

$$R_I^{T|Low} = R_I^{T|High} = \left\{((\theta_{Ij},\gamma_I),\beta_{Ij}) : \gamma_I+\theta_{Ij}\cdot\beta_{Ij} > \frac{1}{2}\right\} \tag{1.13}$$

Note that for each 'type-belief' pair $((\theta_{Ij},\gamma_I),\beta_{Ij})$, the sign of WT is the same in both conditions. Thus, also the complementary set of 'type-belief' pairs consistent with a *R*ational intermediary choosing *Embezzle* is independent from the condition, *i.e.*, $R_I^{E|Low} = R_I^{E|High}$.

Figure 1.3 shows the regions of the intermediary's type space (θ_{Ij},γ_I), where a rational intermediary is predicted to $Embezzle$ in both conditions or $Transfer$ in both conditions for fixed conditional second-order beliefs about *Transfer*.[12] More precisely, for each type (θ_{Ij},γ_I) it is shown the best-reply strategy for $\beta_{Ij} \in \{0,1/3,2/3,1\}$. The four dotted

[11] Note that if the intermediary Transfers he experiences no guilt, and so $B_{Ij} = 0$ in Eq. (1.8).

[12] Recall that, as anticipated above, in our experiment both the donor and the recipient can only hold four possible first-order beliefs about *Transfer*, $\alpha_{jI} \in \{0,1/3,2/3,1\}$ (see Section 3.3). With this, we make the operational assumption that also the intermediary can only hold four possible second-order beliefs, $\beta_{Ij} \in \{0,1/3,2/3,1\}$, on each co-player $j \in \{D,R\}$.

lines indicate types indifferent between *Embezzle* and *Transfer* for each of these β_{Ij}. Thus, *e.g.*, types in the white region *Embezzle* for all the four possible β_{Ij}, while types in the lightest-grey region *Embezzle* for $\beta_{Ij} \in \{0, 1/3, 2/3\}$ and *Transfer* for $\beta_{Ij} = 1$, *i.e.*, they switch from *Embezzle* to *Transfer* for $\beta_{Ij} = 1$.

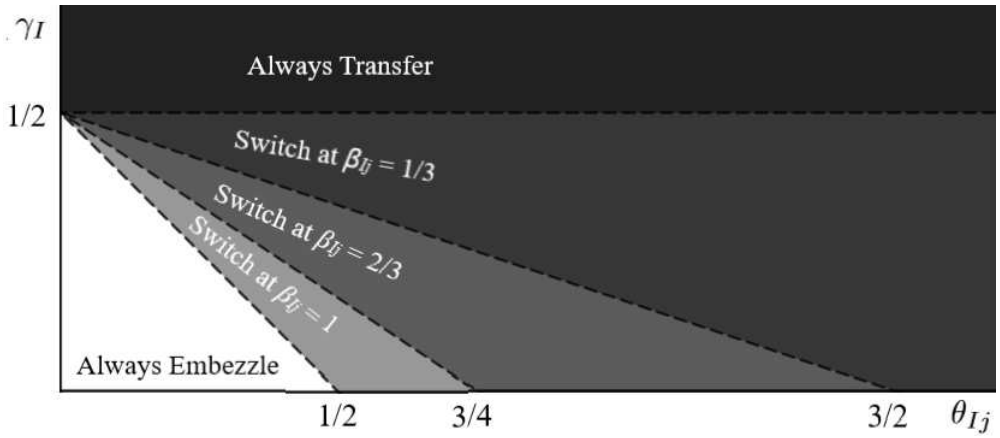


Figure 1.3: Predicted behavior of a rational intermediary for the four possible second-order beliefs $\beta_{Ij} \in \{0, 1/3, 2/3, 1\}$, depending on his guilt type (θ_{Ij}) and altruistic type (γ_I)

A comparative static analysis of WT in Eq. (2.6) and of the four regions of predictions in Figure 3 allows us to elaborate our hypotheses about the intermediary's behavior.

First of all, Figure 1.3 shows that if $\gamma_I > 1/2$, then the intermediary always Transfers, independently from θ_{Ij} and β_{Ij}. In that case, sensitivity to altruism is sufficiently high to prevail over guilt aversion. For $\gamma_I < 1/2$, if the intermediary is guilt-averse ($\theta_{Ij} > 0$), WT in Eq. (2.6) is increasing in β_{Ij}, *i.e.*, $\frac{\delta WT}{\delta \beta_{Ij}} > 0$ independently from the condition Low or High. Therefore, our first hypothesis is about the intermediary's **belief-dependent behavior**:

H.I1 [Choice-Belief Correlation]: For sufficiently low sensitivity to altruism toward the recipient, the frequency of *Transfer* choices by guilt-averse intermediaries increases in

their second-order beliefs about *Transfer*. Intermediaries with sufficiently high sensitivity to altruism choose to *Transfer* regardless of their second-order beliefs.

The second part of H.I1 suggests that the fraction of guilt-averse intermediaries in the sample of experimental participants might be underestimated by only looking at their behavior. In fact, although donor-guilt or recipient-guilt averse, some intermediaries could disclose a belief-independent *Transfer* pattern due to a sufficiently high sensitivity to altruism toward the recipient.

Furthermore, knowing from Eq. (1.13) that $R_I^{s_I|Low} = R_I^{s_I|High}$ for $s_I \in \{T, E\}$, we can elaborate the following hypothesis about the intermediary's **condition-independent behavior**. Note that conditions are manipulated within-subjects, thus we can assume that the distribution of the intermediaries' types is the same across conditions. Belief elicitation in the two conditions will allow us to check their invariance to the condition, that we assume here:

H.I2 [Low *vs.* High Condition on Choice]: Given the same second-order belief about *Transfer* in both conditions, the frequency of *Transfer* choices by intermediaries is the same in the Low and in the High conditions.

Third, WT in Eq. (2.6) is increasing in γ_{Ij}, *i.e.*, $\frac{\delta WT}{\delta \gamma_{Ij}} > 0$, and, for strictly positive second-order beliefs, in θ_{Ij}, *i.e.*, $\frac{\delta WT}{\delta \theta_{Ij}} > 0$, independently from the condition Low or High. Furthermore, Figure 1.3 shows that, fixing $\gamma_I < 1/2$ in the white region and moving horizontally through consecutive increases in θ_{Ij}, the intermediary switches from *Embezzle* to *Transfer* first for $\beta_{Ij} = 1$, then for $\beta_{Ij} = 2/3$, and finally for $\beta_{Ij} = 1/3$. However, even for $\theta_{Ij} \to \infty$, he will never switch from *Embezzle* to *Transfer* for $\beta_{Ij} = 0$. All this is summarized in the following hypothesis about the intermediary's **type-dependent behavior**:

H.I3 [Choice-Type Correlation]: Given the second-order belief, the frequency of *Transfer* choices increases with the altruism sensitivity; it increases with the guilt sensitivity only if the second-order belief is strictly positive. Furthermore, given a sufficiently low altruism sensitivity, the higher the guilt sensitivity, the lower the second-order belief about *Transfer* sufficient to switch from *Embezzle* to *Transfer*.

Note that we derived H.I1, H.I2, and H.I3 without specifying the treatment (Donor or Recipient), and that these hypotheses should hold in both treatments, if no treatment difference in the distribution of intermediaries' psychological types (θ_{Ij}, γ_I) were detected. Altruism being a distributional preference, hence belief-independent, we expect the sensitivity to altruism not to depend on the fact that the intermediary's belief-dependent strategy relies on the donor's or the recipient's first-order beliefs. Conversely, the sensitivity to guilt might depend on the direction, *i.e.*, on whether it is elicited toward the donor or toward the recipient. Indeed, this is one of the two main research objectives of our study. However, absent previous experimental evidence on this issue, we elaborate our last hypothesis on the intermediary's **treatment-independent behavior** assuming the same distribution of guilt types across treatments:

H.I4 [Donor *vs.* Recipient treatment on Choice]: Under the same distribution of sensitivities to guilt and altruism, intermediaries' behavior is the same in both the Donor and the Recipient treatments.

1.3 Experimental Design and Procedures

We now describe in details how the game has been implemented in the laboratory.

1.3.1 Experimental Design

First-Order Belief Elicitation

First, we elicited the players' first-order beliefs about the donors' and the intermediaries' decisions in the game. Intermediaries and recipients had to report their beliefs about the number of donors, out of three donors randomly selected in the session, who choose to give in the Low and in the High conditions that were played within-subjects. Similarly, donors and recipients had to report their beliefs about the number of intermediaries, out of three intermediaries randomly selected in the session, who choose to transfer the donation in full in each condition (conditional on the donor's decision to give). The belief elicitation was incentivized. For each role, one belief was randomly selected at the end of the session and paid €1 if accurate.[13]

Donors' and Intermediaries' Decision-Making

Second, subjects played the Embezzlement Mini-Game. Two treatments of this game were implemented between-subjects: the Donor treatment and the Recipient treatment.[14] Within-subjects, donors made a binary choice between giving a pre-determined fraction of their endowment and keeping their whole endowment, both in the Low and in the High conditions. These two decisions allow us to test whether the giving decision varies with the percentage potentially embezzled by the intermediary as predicted in Hypothesis H.D2.

[13] This incentivization procedure is the easiest to understand for subjects. Nevertheless, we contend that it is not perfectly incentive-compatible for risk-averse recipients who may under-estimate the probability that donors *Give* and that intermediaries *Transfer* to the recipients. However, this concern is hindered both in theory – since there are four possible beliefs, one cannot be perfectly insured against risk – and in practice – we find an insignificant correlation between risk aversion and beliefs (see Table 1.A.11 in Section 1.A).

[14] We used a between-subject design for studying the intermediaries' donor-guilt aversion and recipient-guilt aversion because we were anxious that using a within-subject design would be confusing for the subjects and would require too much concentration.

Then, intermediaries made binary choices between transferring the entirety of the amount given by the donor or transferring only a pre-determined fraction of this donation, both in the Low and in the High conditions. Whether intermediaries started with the Low or with the High condition was determined randomly at the individual level. These decisions were made under the veil of ignorance, *i.e.*, assuming that the donor had chosen to give a positive amount. We used the belief-dependent menu method of Khalmetski et al. (2015). In each condition, in the Donor (Recipient) treatment, intermediaries made four transfer decisions corresponding to the four possible first-order beliefs of the donor (recipient) on the frequency of intermediaries transferring: the donor's (recipient's) beliefs that none, one, two or three out of three intermediaries transfer in full. To facilitate decision-making, these first-order beliefs were presented in a fixed increasing order (see an example of a decision screen in Section 1.C.1).[15]. Although one might argue that responses elicited with this method are "cold", this method offers several advantages. First, it allows us to rule out potential false-consensus effects without raising the issue of strategic reporting and without using deception. The false-consensus effect could be avoided by communicating the donors' (recipients') true beliefs to the intermediaries. However, it requires choosing between two evils: if the donors (recipients) know that their beliefs will be communicated, they are likely to distort them; and if they do not know that their beliefs will be communicated, the design is arguably deceptive. The menu method avoids these drawbacks. Moreover, it allows us to study guilt aversion at the individual level and, hence, to unveil inter-individual differences that are hidden at the aggregate level (Khalmetski et al., 2015).

At the end of the session, the computer program randomly selected either the Low or the High condition. Given that the donor had given a share of his endowment in this

[15]The use of the menu method is frequent in the experimental literature on guilt aversion (Attanasi et al., 2013; Balafoutas and Fornwagner, 2017; Bellemare et al., 2017, 2018; Dhami et al., 2019; Hauge, 2016; Khalmetski et al., 2015)

condition, the program implemented the intermediary's decision corresponding to the actual belief of the donor or of the recipient, depending on the treatment, in this condition. This determined the donor's, the intermediary's and the recipient's payoffs in this part.

Second-Order Belief Elicitation and Social Norms

Third, before subjects received any feedback on payoffs and others' decisions, we elicited the second-order beliefs of the donors and of the intermediaries on the other players' first-order beliefs, both in the Low and in the High conditions. Donors had to guess their intermediary's and their recipient's first-order beliefs on the donors' decisions (four second-order-beliefs in total). Similarly, intermediaries had to guess their donor's and their recipient's first-order beliefs on the intermediaries' decisions (four second-order-beliefs in total). A second-order belief is considered correct if it corresponds to the partner's actual first-order belief.

Moreover, anticipating that behavior in this game may depend on social norms and on the beliefs about others' social norms, we elicited all the subjects' social norms in the session as well as the donors' and the intermediaries' beliefs about their partners' social norms.[16]

The players' social norms were identified, using the Krupka and Weber (2013) procedure, for each donors' and intermediaries' possible decision both in the Low and in the High conditions. In each condition, players had to rate the social appropriateness of each decision on a four-item scale (eight answers in total). An answer is considered correct if it corresponds to the modal answer of the subjects in the same role. Using coordination games among players with the same role to incentivize this procedure allows

[16]Note that d'Adda et al. (2016) found no difference in responses between eliciting normative judgments *à la* Krupka and Weber (2013) before or after playing the main game.

us to identify whether social norms differ across roles. In fact, similarly to Erkut et al. (2015), we found that social norms do not differ across roles in seven out of eight cases (Kruskal-Wallis tests, see Table 1.A.12 in Section 1.A).[17]

Then, donors had to guess their intermediary's and their recipient's ratings of the social appropriateness of the donors' possible decisions (four answers). Similarly, intermediaries had to guess their donor's and their recipient's ratings of the social appropriateness of the intermediaries' possible decisions (four answers). Recipients had no guess to report.

For each subject, we randomly selected one answer among all those provided during this third part. A correct answer paid €1.

Elicitation of Individual Characteristics

Since our model predicts that guilt proneness (Hypothesis H.I3) and altruism (Hypotheses H.D3, H.I3) affect behavior in the game, we elicited the subjects' social preferences by means of several psychological tests. A survey was completed online about a week prior the laboratory session to limit the risk of contamination between this task and the game. Subjects were paid a flat fee of €7 for completing this survey on time and for showing-up at the session in the laboratory.

The survey was composed of four parts (see Section 1.C.2). In the first part, subjects completed the Guilt and Shame Proneness (GASP) questionnaire of Cohen et al. (2011). We were particularly attentive to the Guilt-Negative-Behavior-Evaluation subscale that assesses one's proneness to feel bad about how one acted. The second part was included to control for potentially relevant psychological traits. It corresponds to the Honesty-

[17] Ratings of social appropriateness differ in one instance only: in the Low condition, intermediaries consider that *Embezzle* is less socially appropriate than donors do.

Humility scale extracted from the 100-item HEXACO Personality Inventory – Revised test (Ashton and Lee, 2008). We were interested in the responses to the Fairness subscale that aims at assessing a tendency to avoid dishonesty. The third part consisted of 16 questions from the Self-Reported Altruism Scale (Rushton et al., 1981). Finally, in the fourth part, we collected standard socio-demographic characteristics, including gender, age, professional status, number of past participations in economic experiments, self-reported risk attitudes (using the procedure of Dohmen et al., 2011), and self-reported time preferences (using the procedure of Vischer et al., 2013).

1.3.2 Procedures

The experiment was conducted at GATE-Lab, Lyon, France. It was computerized using the software Z-Tree (Fischbacher, 2007). Subjects were recruited mainly from the undergraduate student population of local business, engineering, and medical schools by email, using the software Hroot (Bock et al., 2014). 369 subjects participated in a total of 19 sessions. 52.72% are females and the average age is 21.85 years (S.D. = 4.54). Table 1.A.1 in Section 1.A summarizes the characteristics of each session.

When subjects registered for the experiment, about a week before the date of the lab session, they were sent an invitation email to complete the online questionnaire. Completing the questionnaire took about 10 minutes. Participants were informed that they would receive their fixed payment of €7 for this task and for showing-up at the laboratory session. Only those who completed the online questionnaire were allowed to participate in the session. In the lab session, at their arrival subjects were randomly assigned to a cubicle after drawing a tag in an opaque bag. The instructions (see Section 1.C.1) were distributed for each part after completion of the previous part. Before the first part, subjects had to answer a comprehension questionnaire. In the first part, subjects

reported their first-order beliefs and the donors made their decisions. In the second part, the intermediaries made their decisions. In the third part, we elicited the subjects' social norms and second-order beliefs.

Each session lasted about 75 minutes. The average earnings were €17.70 (S.D. = 6.19), including the €7 fee for completing the online questionnaire and for showing-up. Earnings were paid in private in a separate room.

1.4 Results

We begin this section by two comments on social norms and beliefs (see summary statistics and significance tests in Table 1.A.2 in Section 1.A). First, *Give* and *Transfer* choices are rated by the participants as significantly more socially appropriate than, respectively, *Keep* and *Embezzle* choices, in both conditions (Wilcoxon signed rank tests, W hereafter, $p < 0.001$). Second, the donors' actual (non-induced) second-order beliefs (SOB, hereafter) are accurate guesses of the intermediaries' and recipients' first-order beliefs (FOB, hereafter) on the frequency of *Give* choices in both conditions (Mann-Whitney rank sum tests, MW hereafter, between SOB and FOB, smallest $p = 0.44$). However, intermediaries tend to overestimate donors' and recipients' FOB on the frequency of *Transfer* choices (MW tests, $p < 0.05$ in three out of four cases).

In the following, we consider first, the donors' behavior (Section 1.4.1) and next, the intermediaries' behavior (Section 1.4.2). For each behavioral hypothesis, we check for treatment differences under the label [Donor *vs.* Recipient treatments]. Except when specified otherwise, the non-parametric tests are two-sided; an independent observation corresponds to a decision (since only one decision per participant is payoff relevant); the results from the two treatments are pooled.

1.4.1 Donors' Behavior

The comparison of our data to the set of 'type-belief' pairs consistent with a *R*ational donor (Eqs. (1.10) and (1.11)) shows that our model captures 92.30% of the observed behavior (see the details and the implications in terms of altruism sensitivity in Table 1.A.4 to Table 1.A.7 in Section 1.A). Overall, 48.78% of donors chose *Give* in the Low condition and 36.56% in the High condition. Figure 1.4 displays, for each condition, the proportion of donors who choose either *Give* or *Keep*, depending on their FOB on the frequency of *Transfer* choices (see also Table 1.A.3 in Section 1.A). The figure illustrates our first two results on the donors' behavior.

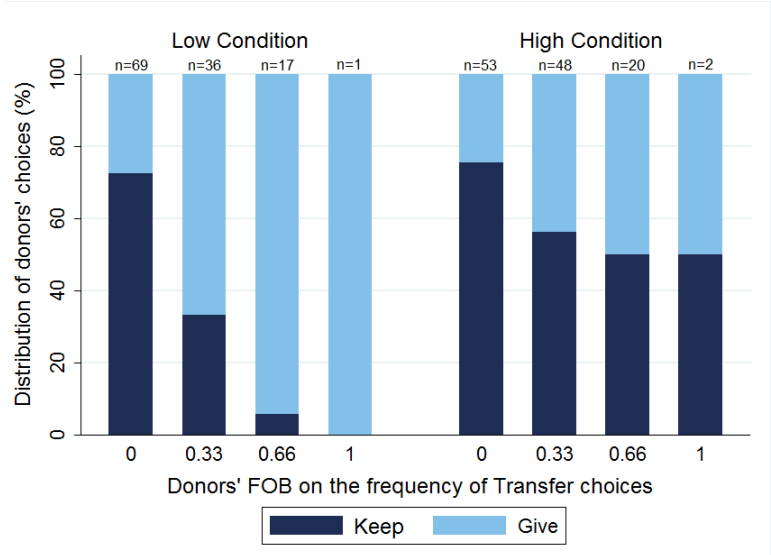


Figure 1.4: Distribution of the donors' choices depending on their first-order beliefs

Result D1 [Choice-Belief Correlation]: The higher the donors' FOB about *Transfer*, the higher the frequency of *Give* choices. This holds in both conditions.[18]

[18] One may suspect that an experimenter demand effect might explain donors' giving despite the sure loss of material payoff entailed by the *Give* choice. However, the detected positive correlation between donors' *Give* choices and their first-order belief of intermediary's *Transfer* choices makes us confident that an experimenter demand effect is not the main driver of donors' behavior.

Support for Result D1: There is a significant positive correlation between the donors' FOB about *Transfer* and their decision to *Give* (Spearman rank correlation, S hereafter, $r_s = 0.35,\ p < 0.001$). When we distinguish between conditions, the correlation in the Low condition (S correlation, $r_s = 0.51,\ p < 0.001$) is significantly higher than in the High condition (S correlation, $r_s = 0.22,\ p < 0.001$) (ZPF statistic, $z = 3.05$, $p < 0.001$).[19]

[Donor *vs.* Recipient treatments]: The correlation between the donors' FOB on the frequency of *Transfer* choices and their decision to *Give* is not significantly different across treatments (Donor treatment: $r_s = 0.44$, $p < 0.001$; Recipient treatment: $r_s = 0.24$, $p < 0.001$; Z test, $z = -1.23$, $p = 0.210$).

Result D2 [High *vs.* Low Condition on Choice]: Controlling for the donors' FOB about *Transfer*, the frequency of *Give* choices is higher in the Low than in the High condition.

Support for Result D2: We use Mc Nemar tests (MN, hereafter) to consider each donor as an independent observation. For a given FOB about *Transfer*, the frequency of *Give* choices is significantly higher in the Low than in the High condition (MN tests; FOB(0): χ^2=4.76, $p = 0.029$; FOB (0.33): χ^2=3.60, $p = 0.057$).[20]

[19]The correlation between the donors' FOB and their decision to *Give* must be regarded with caution. Although belief elicitation was incentivized, it is possible that donors who planned to *Keep* may have underestimated their FOB about *Transfer* choices to justify their selfish choice. To further test H.D1, we consider the donors' rating of the social appropriateness of *Embezzle* as a proxy for their FOB on the frequency of *Transfer* choices because (i) they are significantly correlated (S correlation, $r_s = -0.19$, $p < 0.001$), and (ii) we believe that it is more unlikely that donors used their rating of the social appropriateness of *Embezzle*, rather than their FOB, as a justification of their choice. We replicate the correlation with the ratings of the social appropriateness of *Embezzle* (S correlation, $r_s = -0.20$, $p < 0.001$).

[20]Two donors had a FOB of 0.66 in both conditions and no donor had a FOB of 1 in both conditions.

[Donor *vs.* Recipient treatments]: Even though donors could not know which treatment was implemented when they made their choices, our results differ across treatments. Result D2 is supported in the Donor treatment (MN tests; FOB(0): χ^2=3.57, $p = 0.058$; FOB (0.33): χ^2=3.00, $p = 0.083$) but not in the Recipient Treatment (MN tests; FOB(0): χ^2=1.60, $p = 0.205$; FOB (0.33): χ^2=1.29, $p = 0.252$).

Result D3 [Choice-Type Correlation]: Controlling for the donors' FOB about *Transfer*, the frequency of *Give* choices tends to increase with the donor's sensitivity to altruism. This holds in both conditions.

Support for Result D3: We use the Self-Reported Altruism score (Rushton et al., 1981) as a proxy for our altruism sensitivity parameter. For a given FOB of 0.33, there is a marginally significant positive correlation between the donors' *Give* choices and their altruism score (S correlation, $r_s = 0.42$, $p = 0.081$). However, the correlation is not significant when the given FOB is 0 (S correlation, $r_s = 0.15$, $p = 0.313$). As for Result D2, we cannot test our hypothesis with the other two FOB.

[Donor *vs.* Recipient treatments]: The correlation between the donors' *Give* choices and their altruism score is not different across treatments (Z test, FOB(0): $z = -0.93$, $p = 0.176$; FOB(0.33): $z = 0.61$, $p = 0.270$).

1.4.2 Intermediaries' Behavior

The comparison of our data to the set of 'type-belief' pairs consistent with a *R*ational intermediary (Eq. (1.13)) shows that our model captures 82.93% of the observed behavior: (i) 46.75% of the intermediaries always chose *Embezzle*, (ii) 24.39% switched from

Embezzle to *Transfer* as the induced SOB increases, *i.e.*, exhibiting guilt aversion, and (iii) 11.79% always chose *Transfer*, *i.e.*, exhibiting altruistic preferences prevailing over guilt aversion.[21]

Focusing on behavior consistent with our theoretical predictions, Figure 1.5 presents the distribution of the switching SOB observed in the two treatments and the implications in terms of predicted altruism sensitivity and guilt sensitivity (see Figure 1.3). The distributions of switching SOB do not differ significantly across treatments (Kruskal-Wallis test, $p > 0.10$).

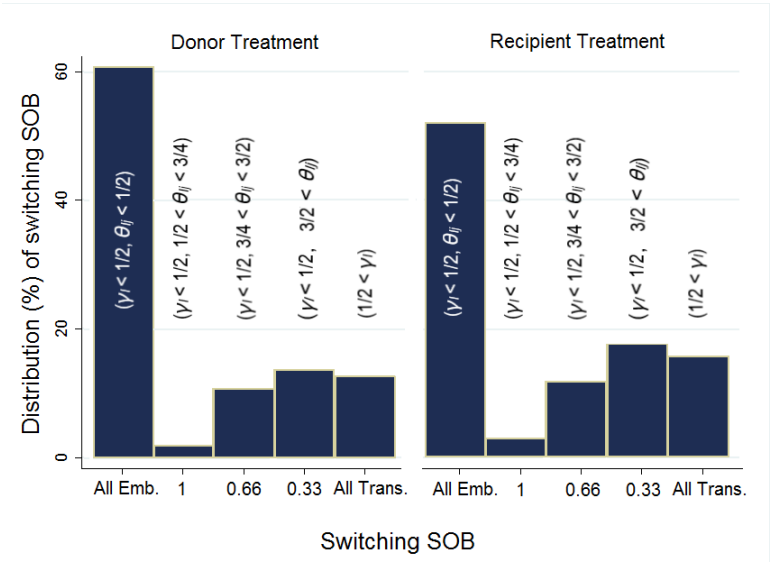


Figure 1.5: Distribution of the intermediaries' switching second-order beliefs

Notes: Subjects who did not behave consistently with our theoretical predictions are excluded from this figure. The figure reads as follows. In the Recipient Treatment, 3% of the intermediaries have a switching SOB of 1, *i.e.*, they chose *Embezzle* for an induced SOB in $\{0; 0.33; 0.66\}$ and *Transfer* for an induced SOB of 1. This behavior is consistent with guilt aversion prevailing over altruism only if $\gamma_I < \frac{1}{2}$ and $\frac{1}{2} < \theta_{Ij} < \frac{3}{4}$ (see Figure 1.3).

[21] Figure 1.3 shows that intermediaries who always choose *Transfer* for any second-order belief have a sensitivity to altruism $\gamma_I > 1/2$, but they can also have a sensitivity to guilt $\theta_{Ij} > 0$. Since we do not know how many of them have $\theta_{Ij} > 0$, the fraction of intermediaries exhibiting guilt aversion might be underestimated in our sample of participants.

The remaining intermediaries behaved as follows: 11.79% of intermediaries switched multiple times between transferring and embezzling, and 5.28% exhibit an inverse switching pattern from transferring to embezzling.

Figure 1.6 displays, for each condition, the proportion of intermediaries who chose either *Transfer* or *Embezzle*, depending on their induced SOB (see also Table 1.A.8 in Section 1.A). The figure illustrates our first two results on the intermediaries' behavior.

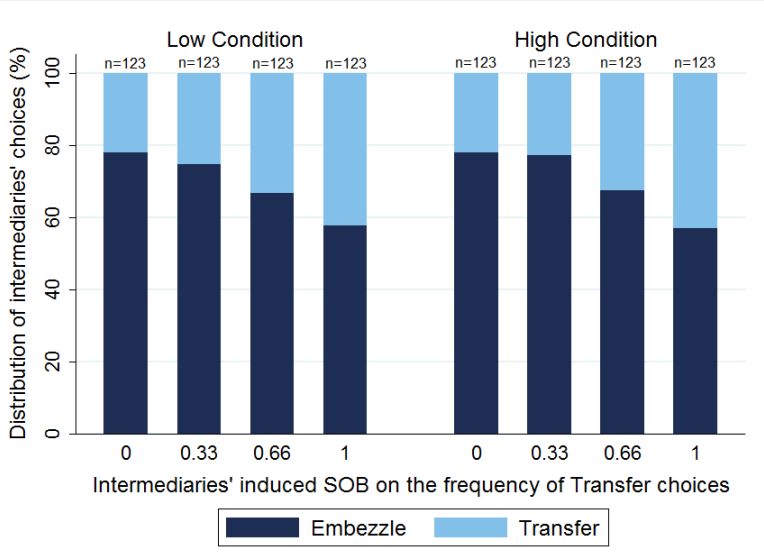


Figure 1.6: Distribution of the intermediaries' choices depending on their induced second-order beliefs

Result I1 [Choice-Belief Correlation]: The higher the intermediaries' induced SOB about *Transfer*, the higher the frequency of *Transfer* choices. This holds in both conditions.

Support for Result I1: There is a significant positive correlation between the intermediaries' induced SOB about *Transfer* and their *Transfer* choices (S correlation, $r_s = 0.15$, $p < 0.001$). The correlation does not vary between conditions (Low: $r_s = 0.17$,

$p < 0.001$; High: $r_s = 0.18$, $p < 0.001$).

Note that, if we exclude the intermediaries who believed that no donor would *Give* in either condition, the correlation increases to $r_s = 0.22$ ($p < 0.001$). Indeed, these excluded intermediaries may suffer from a hypothetical bias, as they are sure that their choices will not be payoff-relevant, rendering the hypothetical decision to embezzle less psychologically costly.

So far, we have conducted the analysis by examining the induced SOB based on the menu method of Khalmetski et al. (2015). If, instead, we use the stated SOB (second-order beliefs reported directly by the subjects in the third part of the experiment), we find that the correlation between the intermediaries' *Transfer* choices and their stated SOB increases to $r_s = 0.27$ ($p < 0.001$). [22] Experiments using stated SOB should not ignore this effect as it leads to an upward-bias measure of the correlation between SOB and choices (see consistent results in Bellemare et al., 2017 and Khalmetski et al., 2015 We also find support for Result I1 using a Logit model with fixed effects (Table 1.A.9 in Section 1.A) and with random effects and individual controls (Table 1.A.10 in Section 1.A).

[Donor *vs.* Recipient treatments]: The correlation between the intermediaries' induced SOB about *Transfer* and their *Transfer* choices does not vary significantly across treatments (Donor treatment: $r_s = 0.14$, $p < 0.001$; Recipient treatment: $r_s = 0.15$, $p < 0.001$; Z test, $z = 0.05$, $p = 0.95$) (see also Table 1.A.9 and Table 1.A.10 in Section 1.A).

Result I2 [High *vs.* Low condition on Choice]: Controlling for the intermediaries' induced SOB about *Transfer*, the frequency of *Transfer* choices is the same in both

[22] We interpret this increase as evidence of a false-consensus effect (Ross et al., 1977; Vanberg, 2008).

conditions.

Support for Result I2: We use MN tests to consider each intermediary as an independent observation. For a given induced SOB, the frequency of *Transfer* choices does not significantly differ across conditions (smallest $p = 0.438$) (see also Table 1.A.10).

[Donor *vs.* Recipient treatments]: We replicate this result when we distinguish between the Donor and the Recipient treatments in seven out of eight cases (MN tests for each induced SOB, smallest $p = 0.256$), with one exception (Recipient treatment when SOB = 0.33: $\chi^2 = 4.50$, $p = 0.033$) (see also Table 1.A.8 in Section 1.A).

Result I3 [Choice-Type Correlation]: The frequency of *Transfer* choices increases (i) for a given second-order belief, with the altruism sensitivity, and (ii) for a given second-order belief, with the guilt sensitivity. Furthermore, the higher the guilt sensitivity, the lower the second-order belief about Transfer sufficient to switch from *Embezzle* to *Transfer*. This holds in both conditions.

Support for Result I3: We consider the Guilt-Negative-Behavior-Evaluation score (Guilt-NBE score, hereafter) elicited in the pre-experimental survey as a proxy for the guilt-sensitivity parameter in our model and the Self-Reported Altruism score (Rushton et al., 1981) as a proxy for the altruism-sensitivity parameter. Table 1.1 presents (i) the correlation between the *Transfer* choices, holding the induced SOB constant, and the Guilt-NBE score, (ii) the correlation between the switching SOB and the Guilt-NBE score, as well as (iii) the correlation between the *Transfer* choices, holding the induced

SOB constant, and the Altruism score. The switching SOB corresponds to the minimum induced SOB sufficient to choose *Transfer* rather than *Embezzle*.[23]

		Donor treatment			Recipient treatment		
		Guilt	Guilt x Hypoth.	Z-stat	Guilt	Guilt x Hypoth.	Z-stat
Transfer	SOB=0	0.10	0.27**	-0.95	0.17	0.23*	-0.34
Transfer	SOB=0.33	0.19	0.39***	-1.18	0.43***	0.46***	-0.20
Transfer	SOB=0.66	0.15	0.47***	-1.93**	0.19*	0.43***	-1.44*
Transfer	SOB=1	0.11	0.49***	-2.29**	0.23*	0.52***	-1.84**
Switching SOB		-0.14	-0.49***	2.07***	-0.10	-0.34***	1.33***
		Altruism	Altruism x Hypoth.	Z-stat	Altruism	Altruism x Hypoth.	Z-stat
Transfer	SOB=0	-0.01	0.28**	-1.60*	0.06	0.14	-0.44
Transfer	SOB=0.33	-0.01	0.35***	-2.02**	0.23***	0.27***	-0.23
Transfer	SOB=0.66	0.02	0.46***	-2.57***	0.03	0.25**	-1.21
Transfer	SOB=1	-0.08	0.44***	-2.97***	0.02	0.39***	-2.11**

Notes: This table presents the coefficients of S correlations between row and column variables. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Rows: "Transfer | SOB = β_{Ij}" represents the total number of *Transfer* choices in both conditions given that the induced SOB was β_{Ij}. "Switching SOB" represents the switching SOB of intermediaries who either always *Embezzle* or are guilt-averse. Columns: Guilt stands for Guilt-NBE score (GASP questionnaire). Altruism stands for Self-Reported Altruism score . Hypoth. stands for a dummy variable that takes value 0 if the intermediary believes that no donor will choose *Give* in either condition, and 1 otherwise. Z-stat stands for the differences between columns measured by Fisher r-to-Z transformations (one-tailed).

Table 1.1: Correlation between the intermediaries' decisions and their Guilt-NBE and Altruism scores

The Guilt-NBE score in itself is only marginally significantly correlated with *Transfer* choices. However, the strength of the correlation improves if we interact this score with a dummy variable that takes value 0 if the intermediary believes that no donor will choose *Give* in either condition, and 1 otherwise (see Table 1.1). The same remark holds for Altruism score, although the improvement is significant only in the Donor Treatment

[23] For an intermediary who always Transfers, the switching SOB is 0; for an intermediary who Embezzles when the induced SOB is 0 and Transfers when the induced SOB is in $\{0.33; 0.66; 1\}$, the switching SOB is 0.33; etc. We cannot compute a switching SOB for intermediaries who exhibited multiples switches or an inverse switching pattern.

(see Table 1.1). This suggests that the Guilt-NBE score and the Self-Reported Altruism score are relevant proxies for, respectively, the guilt-sensitivity and altruism-sensitivity parameters only when intermediaries believe their decision will be implemented with a non-null probability.

[Donor *vs.* Recipient treatments]: The magnitude of this correlation is lower in the Donor treatment than in the Recipient treatment, but not significantly so (Z tests, smallest $p = 0.143$).

Result I4 [Donor *vs.* Recipient]: All our hypotheses hold independently of whether guilt is directed toward the donor or toward the recipient.

Support for Result I4: For each result I1–I3 above, see (the absence of) treatment difference under the label [Donor *vs*. Recipient treatments].

1.4.3 A Structural Estimate of Guilt Sensitivity

Following Bellemare *et al.* (2011), we define a structural econometric model to estimate the intermediaries' average guilt-sensitivity parameter, θ_{Ij}, toward the donor ($j = D$, in the Donor treatment) and the recipient ($j = R$, in the Recipient treatment).[24] Given the treatment, for each α_{jI} and each condition (eight cases per intermediary), intermediaries choose s_I (*Transfer* or *Embezzle*) to maximize their utility after *Give*, as defined by Equation (2.9) (Random Utility Model). In this equation, λ is the noise parameter that we

[24] Recall that, in our theoretical model, we also introduce a parameter which represents the altruism sensitivity, γ_I (see Eq. (1.7)). However, the second component of the intermediary's feeling of altruism A_{IR}, *i.e.*, the recipient's received amount $r(s_I)$, is colinear with the intermediary's material payoff (by construction of the Mini-Games): $r(s_I) = 2 \cdot [25 - (M_I(s_I) - 80)]$ in both conditions. Therefore, we cannot estimate the three coefficients (γ_I, θ_{Ij}, and the coefficient corresponding to M_I) of our theoretical utility function while estimating the noise parameter of our random utility model (Eq. (2.9)). We renounce to estimate γ_I.

estimate, and U_I is defined following our modeling of guilt aversion toward the donor or the recipient (Eq. 1.6): $U_I(\theta_{Ij}, s_I | s_D = G, \alpha_{jI}) = 1 \cdot M_I(G, s_I) - \theta_{Ij} \cdot \max\{0, \mathbb{E}_j[r(s_I)|s_D = G] - r(s_I)\}$ in the Low and High condition, for $\alpha_{jI} \in \{0, 1/3, 2/3, 1\}$, and $j \in \{D, R\}$:

$$V_I(\theta_{Ij}, \lambda, s_I) \quad = \quad U_I(\theta_{Ij}, s_I) + \lambda \cdot \varepsilon_I(s_I) \tag{1.14}$$

We used a conditional Logit model to estimate θ_{Ij}, the coefficient corresponding to the guilt sensitivity parameter, and λ, the noise parameter, while fixing to 1 the coefficient corresponding to the intermediary's own material payoff. Table 2.5.2 reports the results of these estimates.

	All treatments	Donor treatment	Recipient treatment	Subjects with Hypothetical Bias Excluded
θ_{Ij}	-0.37***	-0.34***	-0.41***	-0.61***
	(0.03)	(0.04)	(0.04)	(0.04)
λ	6.48***	5.80***	7.32***	8.30
	(0.42)	(0.51)	(0.74)	(0.79)
N	123	62	61	83

Notes: Standard errors in parentheses; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 1.2: Structural estimates of the guilt-sensitivity parameter

The results reported in Table 2.5.2 show that the average intermediary is willing to pay 0.37 ECU to avoid letting down another player by 1 ECU (difference between expectations and actual outcome). When we exclude intermediaries who believed that no donor chose to *Give* (those intermediaries who are potentially subject to a hypothetical bias), the estimated guilt-sensitivity parameter increases up to 0.61.

[Donor *vs.* Recipient treatments]: Although intermediaries seem to be slightly more sensitive to guilt toward the recipient than toward the donor (+20%), the difference is not significant (Z test, $z = -1.09$, $p = 0.13$; see Paternoster et al. (1998)).

1.5 Discussion and Conclusion

In this study we investigated theoretically and experimentally the role of guilt aversion in the behavior of intermediaries confronted with an opportunity to embezzle a donation. Using psychological game theory, our aim was to determine (i) whether others' expectations influence the decision to embezzle, and (ii) whether the impact of others' expectations on behavior differs if others are the donors or the potential recipients of the donation. Extending Battigalli and Dufwenberg (2007) model to capture guilt aversion toward the donor and documenting its existence and features in a laboratory experiment are our two original contributions. Indeed, we have modeled a new direction of guilt whose existence was not documented yet: guilt directed toward a player whose payoffs cannot be affected by the agent's decision. The recent experimental literature on guilt aversion has often pursued three separate objectives: measuring the prevalence of guilt aversion in the population and its magnitude, and identifying a survey-based measure of guilt aversion. We are the first to address these three questions in a single paper.

We find that (i) on average, about 25 % of the intermediaries are affected by others' expectations in the way predicted by our guilt-aversion model, and the proportion of guilt-averse intermediaries is not affected by the direction of guilt; (ii) on average, an intermediary is willing to pay 0.37 ECU not to let down another player by 1 ECU, and the intensity of the structurally estimated guilt-sensitivity parameter is not significantly different when the intermediary is confronted with the recipient's expectations (0.41) compared to the donor's expectations (0.34). Thus, guilt aversion has the same effect on intermediaries, regardless of whether the intermediary considers a person that may be financially harmed by his decision or a person that he may betray but without any monetary consequences.

Our results contribute to the recent strand of the literature aiming at estimating the proportion of guilt-averse individuals in the population — a literature so far limited to Dictator games (see Table 1.B.1 in Section 1.B). Our structural estimates of guilt sensitivity are in the same range of values as those obtained by Bellemare et al. (2011, 2018) through structural estimations (see Table 1.B.2 in Section 1.B). Finally, we report a significant positive correlation between the intermediaries' switching second-order beliefs and their Guilt-Negative-Behavior-Evaluation score, but only when intermediaries believe that their decision will be implemented with a non-null probability that they are not playing hypothetically. This finding contributes to the small literature trying to identify the link between survey-based measures of guilt and experimental decisions (see Table 1.B.3 in Section 1.B). Overall, this calls for more research on the nature of the emotions embedded in Battigalli and Dufwenberg (2007) model of guilt-aversion.

These findings highlight that psychological game theory can contribute usefully to the renewal of the analysis of dishonesty by a better understanding of the moral costs of unethical behavior. We measured guilt aversion toward the donor and toward the recipient in two separate treatments. A straightforward extension would be to test a treatment in which intermediaries would be informed about both donors' and recipients' expectations. This would lead to a complex design, though. By enlarging the perspective to a dynamic setting, we could also contribute to explain the emergence of a vicious circle of corrupt norms. If donors or recipients expect a high level of embezzlement in a group, intermediaries can embezzle without feeling guilty, which in turn increases the expectations of embezzlement.

If the results on intermediaries' guilt aversion in the lab hold in the field, anti-corruption policies could publicize the high expectations of donors and recipients to the interme-

diaries. Public campaigns of information (Reinikka and Svensson, 2011) or framing manipulations (Ockenfels and Werner, 2014) usually focus on the potential recipients' expectations. Policies should also consider the sensitiveness of intermediaries to the donors' expectations (see also the literature on trust-responsiveness, *e.g.*, Bacharach et al., 2007; Guerra and Zizzo, 2004). But of course, identifying guilt aversion in the lab does not prove that it exists to the same extent in the field. In the field there may be an asymmetry in guilt aversion because the hierarchy of status or power adds to the inequality of payoffs that we introduced in our experiment (for example, donors are sometimes a corrupt and exploitative government dealing with other people's money; thus, guilt toward the recipients may be much stronger than toward the donor). Note that previous studies on bribery found no difference in behavior in the field and in the lab (Armantier and Boly, 2013) and that dishonesty in the lab correlates with dishonesty of the same individuals in the field (Dai et al., 2018). Future research should usefully test the qualitative and quantitative external validity of our results. A major challenge, though, will be to measure beliefs in the field.

Bibliography

Abbink, K. and Serra, D. (2012). Chapter 4 anticorruption policies: Lessons from the lab. In *New advances in experimental research on corruption*, pages 77–115. Emerald Group Publishing Limited.

Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, 87(4):1115–1153.

Armantier, O. and Boly, A. (2013). Comparing corruption in the laboratory and in the field in burkina faso and in canada. *The Economic Journal*, 123(573):1168–1187.

Ashton, M. C. and Lee, K. (2008). The prediction of honesty–humility-related criteria by the hexaco and five-factor models of personality. *Journal of Research in Personality*, 42(5):1216–1228.

Attanasi, G., Battigalli, P., and Manzoni, E. (2016). Incomplete-information models of guilt aversion in the trust game. *Management Science*, 62(3):648–667.

Attanasi, G., Battigalli, P., Manzoni, E., and Nagel, R. (2018). Belief-dependent preferences and reputation: Experimental analysis of a repeated trust game. *Journal of Economic Behavior & Organization*.

Attanasi, G., Battigalli, P., Nagel, R., et al. (2013). Disclosure of belief-dependent preferences in the trust game. *IGIER Working Papers*, 506.

Bacharach, M., Guerra, G., and Zizzo, D. J. (2007). The self-fulfilling property of trust: An experimental study. *Theory and Decision*, 63(4):349–388.

Bahr, G. and Requate, T. (2014). Reciprocity and giving in a consecutive three-person dictator game with social interaction. *German Economic Review*, 15(3):374–392.

Balafoutas, L. (2011). Public beliefs and corruption in a repeated psychological game. *Journal of Economic Behavior & Organization*, 78(1-2):51–59.

Balafoutas, L. and Fornwagner, H. (2017). The limits of guilt. *Journal of the Economic Science Association*, 3(2):137–148.

Barr, A., Lindelow, M., and Serneels, P. (2009). Corruption in public service delivery: An experimental analysis. *Journal of Economic Behavior & Organization*, 72(1):225–239.

Battigalli, P., Corrao, R., and Dufwenberg, M. (2019a). Incorporating belief-dependent motivation in games. *Journal of Economic Behavior & Organization*, 185-218:185–218.

Battigalli, P., Corrao, R., and Sanna, F. (2019b). Epistemic game theory without types structures: An application to psychological games. *IGIER Working Papers*, 641.

Battigalli, P. and Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, 97(2):170–176.

Battigalli, P. and Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory*, 144(1):1–35.

Becker, G. S. and Stigler, G. J. (1974). Law enforcement, malfeasance, and compensation of enforcers. *The Journal of Legal Studies*, 3(1):1–18.

Beekman, G., Bulte, E., and Nillesen, E. (2014). Corruption, investments and contributions to public goods: Experimental evidence from rural liberia. *Journal of public economics*, 115:37–47.

Bellemare, C., Sebald, A., and Strobel, M. (2011). Measuring the willingness to pay to avoid guilt: estimation using equilibrium and stated belief models. *Journal of Applied Econometrics*, 26(3):437–453.

Bellemare, C., Sebald, A., and Suetens, S. (2017). A note on testing guilt aversion. *Games and Economic Behavior*, 102:233–239.

Bellemare, C., Sebald, A., and Suetens, S. (2018). Heterogeneous guilt sensitivities and incentive effects. *Experimental Economics*, 21(2):316–336.

Bock, O., Baetge, I., and Nicklisch, A. (2014). hroot: Hamburg registration and organization online tool. *European Economic Review*, 71:117–120.

Bolton, G. E. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American economic review*, 90(1):166–193.

Boly, A., Gillanders, R., and Miettinen, T. (2016). Deterrence, peer effect, and legitimacy in anti-corruption policy-making: An experimental analysis. *WIDER Working Paper*.

Canagarajah, S. and Ye, X. (2001). *Public health and education spending in Ghana in 1992–98: Issues of equity and efficiency*. World Bank Publications.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6):1579–1601.

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

Chlaß, N., Gangadharan, L., and Jones, K. (2015). Charitable giving and intermediation. *Jena Economic Research Papers*.

Cohen, T. R., Wolf, S. T., Panter, A. T., and Insko, C. A. (2011). Introducing the gasp scale: a new measure of guilt and shame proneness. *Journal of personality and social psychology*, 100(5):947.

d'Adda, G., Drouvelis, M., and Nosenzo, D. (2016). Norm elicitation in within-subject designs: Testing for order effects. *Journal of Behavioral and Experimental Economics*, 62:1–7.

Dai, Z., Galeotti, F., and Villeval, M. C. (2018). Cheating in the lab predicts fraud in the field: An experiment in public transportation. *Management Science*, 64(3):1081–1100.

Dhami, S., Wei, M., and al Nowaihi, A. (2019). Public goods games and psychological utility: Theory and evidence. *Journal of Economic Behavior & Organization*, 167:361–390.

Di Falco, S., Magdalou, B., Masclet, D., Villeval, M. C., and Willinger, M. (2016). Can transparency of information reduce embezzlement? experimental evidence from tanzania. *GATE Working Papers*, (1618).

Di Tella, R. and Schargrodsky, E. (2003). The role of wages and auditing during a crackdown on corruption in the city of buenos aires. *The Journal of Law and Economics*, 46(1):269–292.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3):522–550.

Drugov, M., Hamman, J., and Serra, D. (2014). Intermediaries in corruption: an experiment. *Experimental Economics*, 17(1):78–99.

Dufwenberg, M. and Dufwenberg, M. A. (2018). Lies in disguise–a theoretical analysis of cheating. *Journal of Economic Theory*, 175:248–264.

Erkut, H., Nosenzo, D., and Sefton, M. (2015). Identifying social norms using coordination games: Spectators vs. stakeholders. *Economics Letters*, 130:28–31.

Fan, C. S., Lin, C., and Treisman, D. (2009). Political decentralization and corruption: Evidence from around the world. *Journal of Public Economics*, 93(1-2):14–34.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3):817–868.

Ferraz, C., Finan, F., and Moreira, D. B. (2012). Corrupting learning: Evidence from missing federal education funds in brazil. *Journal of Public Economics*, 96(9-10):712–726.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, 10(2):171–178.

Fischbacher, U. and Föllmi-Heusi, F. (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association*, 11(3):525–547.

Geanakoplos, J., Pearce, D., and Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and economic Behavior*, 1(1):60–79.

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, 95(1):384–394.

Guerra, G. and Zizzo, D. J. (2004). Trust responsiveness and beliefs. *Journal of Economic Behavior & Organization*, 55(1):25–30.

Hauge, K. E. (2016). Generosity and guilt: The role of beliefs and moral standards of others. *Journal of Economic Psychology*, 54:35–43.

Kajackaite, A. and Gneezy, U. (2017). Incentives and cheating. *Games and Economic Behavior*, 102:433–444.

Khalmetski, K., Ockenfels, A., and Werner, P. (2015). Surprising gifts: Theory and laboratory evidence. *Journal of Economic Theory*, 159:163–208.

Köbis, N. C., van Prooijen, J.-W., Righetti, F., and Van Lange, P. A. (2016). Prospection in individual and interpersonal corruption dilemmas. *Review of General Psychology*, 20(1):71–85.

Krupka, E. L. and Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3):495–524.

Mazar, N., Amir, O., and Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of marketing research*, 45(6):633–644.

Ockenfels, A. and Werner, P. (2014). Scale manipulation in dictator games. *Journal of Economic Behavior & Organization*, 97:138–142.

Olken, B. A. (2007). Monitoring corruption: evidence from a field experiment in indonesia. *Journal of political Economy*, 115(2):200–249.

Olken, B. A. and Pande, R. (2012). *Accessed corruption in developing countries*. MIT Annual Reviews.

Paternoster, R., Brame, R., Mazerolle, P., and Piquero, A. (1998). Using the correct statistical test for the equality of regression coefficients. *Criminology*, 36(4):859–866.

Reinikka, R. and Svensson, J. (2004). Local capture: evidence from a central government transfer program in uganda. *The quarterly journal of economics*, 119(2):679–705.

Reinikka, R. and Svensson, J. (2011). The power of information in public services: Evidence from education in uganda. *Journal of Public Economics*, 95(7-8):956–966.

Ross, L., Greene, D., and House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology*, 13(3):279–301.

Rushton, J. P., Chrisjohn, R. D., and Fekken, G. C. (1981). The altruistic personality and the self-report altruism scale. *Personality and individual differences*, 2(4):293–302.

Tangney, J. and Fisher, K. (1995). *Self-conscious emotions: the psychology of shame, guilt and pride*. New York: The Guilford Press.

Vanberg, C. (2008). Why do people keep their promises? an experimental test of two explanations 1. *Econometrica*, 76(6):1467–1480.

Vischer, T., Dohmen, T., Falk, A., Huffman, D., Schupp, J., Sunde, U., and Wagner, G. G. (2013). Validating an ultra-short survey measure of patience. *Economics Letters*, 120(2):142–145.

Appendix

1.A Additional Results

Session (#)	Participants (n)	Age (mean)	Women (%)	Previous Exp. (mean)	Economics Stud. (%)
			Donor Treatment		
5	18	21.27	50.00	1.44	66.67
6	18	21.00	72.22	0.22	33.33
7	15	24.20	53.33	1.53	33.33
8	21	22.00	90.48	0.80	28.57
9	21	21.28	42.86	1.80	71.43
12	24	23.25	50.00	1.33	45.83
13	18	20.66	55.56	1.50	61.11
15	21	21.19	47.82	1.00	38.10
16	12	21.00	50.00	2.50	50.00
18	15	22.40	46.67	1.86	33.33
Sub-total	183	21.83	56.28	1.34	46.45
			Recipient Treatment		
1	18	21.77	22.22	1.16	77.78
2	21	19.76	61.90	0.90	57.14
3	15	20.93	40.00	0.26	80.00
4	21	20.85	57.14	1.14	52.38
10	18	22.88	72.22	1.16	61.11
11	27	22.30	50.00	1.96	62.96
14	24	21.50	37.50	2.20	54.17
17	27	24.59	40.74	2.70	55.56
19	15	21.00	66.67	2.46	46.67
Sub-total	186	21.87	49.18	1.63	60.22
Treatment Diff.		No[1]	No[2]	No[1]	Yes[2]***
Total	369	21.85	52.72	1.49	53.39

Notes: [1] Mann-Whitney ranks sum tests; [2] Fisher exact test

Table 1.A.1: Summary statistics of participants per session

	Low	z-stat	High	z-stat
On the donors' behavior				
Intermediaries' FOB on the frequency of Give choices [a]	0.39	0.42	0.37	-0.21
Donors' SOB on intermediaries' FOB [a]	0.40		0.35	
Recipients' FOB on the frequency of Give choices [a]	0.40	-0.76	0.36	-0.30
Donors' SOB on recipients' FOB [a]	0.37		0.34	
Social Norm on Give [b]	0.88	-16.10***	0.84	-15.78***
Social Norm on Keep [b]	-0.48		-0.43	
On the intermediaries' behavior				
Donors' FOB on the frequency of Transfer choices [a]	0.20	-4.72***	0.25	-2.14**
Intermediaries' SOB on donors' FOB [a]	0.36		0.27	
Recipients' FOB on the frequency of Transfer choices [a]	0.21	3.15**	0.27	1.24
Intermediaries' SOB on recipients' FOB [a]	0.30		0.29	
Social Norm on Transfer [b]	0.89	-14.45***	0.90	-15.74***
Social Norm on Embezzle [b]	0.19		-0.18	

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

[a] Average beliefs on the frequency of choices are rated on scale from 0 (never) to 1 (always). Differences between FOB and SOB are measured by Mann-Whitney rank sum tests.

[b] Average social norms are rated on a scale from -1 (very socially inappropriate) to 1 (very socially appropriate). Differences between social norms are measured by Wilcoxon signed rank tests.

Table 1.A.2: Summary statistics on beliefs and social norms

		Low Condition		High Condition	
		%	n	%	n
Give	FOB=0	27.54%	19	24.53%	13
Give	FOB=0.33	66.67%	24	43.75%	21
Give	FOB=0.66	94.12%	16	50.00%	10
Give	FOB=1	100%	1	50.00%	1

Notes: For each condition, a donor makes one choice given his FOB, e.g., in the Low condition, among the donors whose FOB was 0.33, 66.67% chose *Give*.

Table 1.A.3: Donors' *Give* choices for a given FOB on the frequency of *Transfer* choices

(*Keep*, *Keep*)	FOB = 0		FOB = 0.33		FOB = 0.66		FOB = 1	
	Prediction	n	Prediction	n	Prediction	n	Prediction	n
FOB = 0	$\gamma_D < \frac{5}{4}$	34	$\gamma_D < \frac{15}{14}$	10	$\gamma_D < \frac{15}{22}$	1	$\gamma_D < \frac{1}{2}$	0
FOB = 0.33	$\gamma_D < \frac{5}{2}$	4	$\gamma_D < \frac{5}{6}$	4	$\gamma_D < \frac{15}{22}$	3	$\gamma_D < \frac{1}{2}$	0
FOB = 0.66	$\gamma_D < \frac{5}{8}$	0	$\gamma_D < \frac{5}{8}$	1	$\gamma_D < \frac{5}{8}$	0	$\gamma_D < \frac{1}{2}$	0
FOB = 1	$\gamma_D < \frac{1}{2}$	0	$\gamma_D < \frac{1}{2}$	0	$\gamma_D < \frac{1}{2}$	0	$\gamma_D < \frac{1}{2}$	0

Notes: The table reads as follows. 10 donors chose to *Keep* in the Low condition and to *Keep* in the High condition while having a FOB of 0 about *Transfer* choices in the Low condition and a FOB of 0.33 about *Transfer* choices in the High condition. This behavior is consistent with our theoretical predictions only if $\gamma_D < \frac{15}{14}$.

Table 1.A.4: Matching the donors' behavior to our predictions - (*Keep*, *Keep*)

(*Give*, *Keep*)	FOB = 0		FOB = 0.33		FOB = 0.66		FOB = 1	
	Prediction	n	Prediction	n	Prediction	n	Prediction	n
FOB = 0	$\frac{5}{4} < \gamma_D < \frac{5}{2}$	0	No	1	No	0	No	0
FOB = 0.33	$\frac{5}{2} < \gamma_D < \frac{5}{2}$	2	$\frac{5}{6} < \gamma_D < \frac{15}{14}$	7	No	5	No	0
FOB = 0.66	$\frac{5}{8} < \gamma_D < \frac{5}{2}$	0	$\frac{5}{8} < \gamma_D < \frac{15}{14}$	4	$\frac{5}{8} < \gamma_D < \frac{15}{22}$	1	No	1
FOB = 1	$\frac{1}{2} < \gamma_D < \frac{5}{2}$	0	$\frac{1}{2} < \gamma_D < \frac{15}{22}$	0	$\frac{1}{2} < \gamma_D < \frac{15}{14}$	0	No	0

Notes: the table reads as in Table 1.A.4. "No" means that there exists no value of γ_D leading to a (*Give*,*Keep*) prediction for the specific pair of beliefs in the two conditions.

Table 1.A.5: Matching the donors' behavior to our predictions - (*Give*, *Keep*)

(*Give*, *Give*)	FOB = 0		FOB = 0.33		FOB = 0.66		FOB = 1	
	Prediction	n	Prediction	n	Prediction	n	Prediction	n
FOB = 0	$\frac{5}{2} < \gamma_D$	11	$\frac{5}{4} < \gamma_D$	4	$\frac{5}{4} < \gamma_D$	3	$\frac{5}{4} < \gamma_D$	0
FOB = 0.33	$\frac{5}{2} < \gamma_D$	0	$\frac{15}{14} < \gamma_D$	6	$\frac{5}{6} < \gamma_D$	4	$\frac{5}{6} < \gamma_D$	0
FOB = 0.66	$\frac{5}{2} < \gamma_D$	1	$\frac{15}{14} < \gamma_D$	7	$\frac{15}{22} < \gamma_D$	1	$\frac{5}{8} < \gamma_D$	1
FOB = 1	$\frac{5}{2} < \gamma_D$	0	$\frac{15}{14} < \gamma_D$	0	$\frac{15}{22} < \gamma_D$	1	$\frac{1}{2} < \gamma_D$	0

Notes: the table reads as in Table 1.A.4. "No" means that there exists no value of γ_D leading to a (*Give*,*Give*) prediction for the specific pair of beliefs in the two conditions.

Table 1.A.6: Matching the donors' behavior to our predictions - (*Give*, *Give*)

(*Keep*, *Give*)	FOB = 0		FOB = 0.33		FOB = 0.66		FOB = 1	
	Prediction	n	Prediction	n	Prediction	n	Prediction	n
FOB = 0	No	1	$\frac{15}{14} < \gamma_D < \frac{5}{4}$	3	$\frac{15}{22} < \gamma_D < \frac{5}{4}$	1	$\frac{1}{2} < \gamma_D < \frac{5}{4}$	0
FOB = 0.33	No	0	No	1	$\frac{15}{22} < \gamma_D < \frac{5}{6}$	0	$\frac{1}{2} < \gamma_D < \frac{5}{6}$	0
FOB = 0.66	No	0	No	0	No	0	$\frac{1}{2} < \gamma_D < \frac{5}{8}$	0
FOB = 1	No	0	No	0	No	0	No	0

Notes: the table reads as in Table 1.A.4. "No" means that there exists no value of γ_D leading to a (*Keep*,*Give*) prediction for the specific pair of beliefs in the two conditions.

Table 1.A.7: Matching the donors' behavior to our predictions - (*Keep*, *Give*)

		Low Condition		High Condition	
		%	n	%	n
Transfer	SOB=0	21.95%	27	21.95%	27
Transfer	SOB=0.33	25.20%	31	27.76%	28
Transfer	SOB=0.66	33.33%	41	32.52%	40
Transfer	SOB=1	42.28%	52	43.09%	53

Notes: For each condition, an intermediary makes four choices given each induced SOB, e.g., in the Low condition, when the induced SOB was 0.33, 25.20% of intermediaries chose *Transfer*.

Table 1.A.8: Intermediaries' *Transfer* choices for a given induced SOB

	All treatments	Donor treatment	Recipient treatment	Hypothetical Bias Excluded	All treatments
Induced SOB	0.67***	0.62***	0.75***	0.81***	
	(0.09)	(0.12)	(0.14)	(0.11)	
Low Condition	0.06	0.21	0.13	0.10	-0.24
	(0.20)	(0.26)	(0.30)	(0.22)	(0.52)
Stated SOB					1.27**
					(0.63)
# Observations	472	256	216	400	42
# Participants	59	32	27	50	21

Notes: In the first four columns, the dependent variable is the decision to *Transfer* made for a given induced SOB (using the menu method). In the last column, the dependent variable is the decision to *Transfer* made when the induced SOB corresponded to the stated SOB (the SOB reported directly by the intermediaries in the third part of the experiment). Standard errors in parentheses; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 1.A.9: Regression on the decision to Transfer (Logit model, fixed effects)

	All treatments	Donor treatment	Recipient treatment	Hypothetical Bias Excluded	All treatments
Induced SOB	0.75***	0.74***	0.76***	0.90***	
	(0.10)	(0.13)	(0.15)	(0.11)	
Low Condition	0.08	0.27	-0.13	0.13	-0.04
	(0.20)	(0.28)	(0.30)	(0.23)	(0.47)
Donor Treatment	-0.94			-0.54	-0.93
	(0.60)			(0.61)	(0.77)
Stated SOB					2.16***
					(0.54)
Individual Controls	Yes	Yes	Yes	Yes	Yes
# Observations	876	488	488	656	244
# Participants	122	61	61	82	122

Notes: In the four first columns, the dependent variable is the decision to *Transfer* made for a given induced SOB. In the last column, the dependent variable is the decision to *Transfer* made when the induced SOB corresponded to the stated SOB. Individual controls are: age, gender, guilt-NBE score, fairness score. Standard errors in parentheses; * p $<$ 0.1, ** p $<$ 0.05, *** p $<$ 0.01

Table 1.A.10: Regression on the decision to Transfer (Logit model, random effects)

Correlation between ...	Risk-Aversion
FOB on Donors' Behavior (Low condition)	0.06
FOB on Donors' Behavior (High condition)	0.09
FOB on Intermediaries' Behavior (Low condition)	0.04
FOB on Intermediaries' Behavior (High condition)	-0.02

N=123; Standard errors in parentheses; * $p < 0.1$, ** $p <$0.05, *** $p <$0.01

Table 1.A.11: Correlation between recipients' beliefs and recipients' risk aversion

	Low condition (χ^2)	High condition (χ^2)
Social Norm on Give	4.17	2.59
Social Norm on Keep	3.61	2.39
Social Norm on Transfer	0.21	1.75
Social Norm on Embezzle	6.89**	0.97

Notes: Kruskal-Wallis tests. Standard errors in parentheses; * $p <$0.1, ** $p <$0.05, *** $p <$ 0.01.
[a]: The mean of the intermediaries (0.11) is smaller than the mean of the donors (0.28) (t-test, $p <$0.05).

Table 1.A.12: Difference in social norms distributions across roles

1.B Previous Literature

Study	Game	%	N
Khalmetski et al. (2015)	Dictator	37%	191
Balafoutas and Fornwagner (2017)	Dictator	18%	108
Bellemare et al. (2018)	Dictator	$\approx$ 65%	140
Our results	Embezzlement	25%	123

Table 1.B.1: Previous estimations of the proportion of guilt-averse individuals

Study	Game	Estimation	Treatment	θ_i	N
Bellemare et al. (2011)	Proposal and Response	Structural	Dictators' SOB	0.4	1078
			Recipients' FOB	0.8	540
Bellemare et al. (2018)	Dictator	Structural	Stake-independent	0.1	84
			Low Stakes	0.4	56
			Medium Stakes	0.6	56
			High Stakes	1	56
Patel and Smith (2019)	Participation	Equilibrium		0.1	111
Peeters and Vorsatz (2018)	Prisonner Dilemma	Equilibrium	Baseline	2.3	90
			Tempting to coop.	1.8	92
			Tempting to def.	2.5	96
		Hypothetical BDM	Baseline.	3.1	90
			Tempting to coop.	2.1	92
			Tempting to def.	3.5	96
Our results	Embezzlement	Structural	Toward Donor	0.34	61
			Toward Recipient	0.41	62

Table 1.B.2: Previous estimations of the guilt-sensitivity parameter

		Correlation between ...			
Study	Game	Trait	Behavior	$p < 0.1$	N
Bracht and Regner (2013)	Trust	Guilt-NBE	Pro-social choice	Yes	192
Regner and Harth (2014)	Trust	Moulton's [a]	Pro-social choice	Yes	127
Peeters and Vorsatz (2018)	Prisonner Dilemma	Guilt-NBE	Estimated θ	No	68
Our results	Embezzlement	Guilt-NBE	Pro-social choice	Yes/No	123
			Switching SOB	Yes/No	

Notes: [a] Regner and Harth (2014) used a one question out of the three included in the original measure of Moulton et al. (1966): "How easy is it for something to make you feel guilty? (1) very easy, (2) easy, (3) difficult, (4) very difficult".

Table 1.B.3: Previous correlation of personality traits and behavioral outcomes

1.C Instructions

1.C.1 Instructions for the Lab Experiment [Translated from French]

OVERVIEW OF THE SESSION

Thank you for participating in this experimental session on decision-making. During this session, you can earn money. The amount of your earnings depends both on your decisions and on other participants' decisions. At the end of the session, you will receive your earnings in cash, in a separate room to ensure the confidentiality of your earnings. The earnings you will receive include:

- your earnings from today's experimental session
- a €7 fee for having completed the online questionnaire and for showing-up on time

During the session, we will sometimes use ECU (Experimental Currency Units). The conversion rate from ECU into Euro is the following: 10 ECU = €1.2.

Please turn off your phone. During the session, any communication with other participants is forbidden. If you have any questions, raise your hand or press the red button on the side of your desk. We will come answer to your questions in private.

At the beginning of the session, the program will form groups of three participants. You will never know the identity of the other two members of your group, and they will never know your identity. All your decisions and earnings are anonymous.

In each group, participants have a different role. There is:

- a donor
- an intermediary
- a recipient

Your screen will indicate your role when the session begins and you will keep the same role throughout the session.

There are two possible situations: situation A and situation B. You will take your decisions in both situations. At every moment, the situation in which you are will always be displayed on the screen.

Short description of the roles

ROLE OF THE DONOR

The donor receives an initial endowment of 150 ECU.

The donor's task is to choose how many ECU to give to the recipient.

For each situation, the donor decides either:

- to give 25 ECU to the recipient
- or to give 0 ECU to the recipient

Regardless of the situation, his/her payoff is equal to: 150 ECU – the ECU given.

Important: The donor cannot give ECU directly to the recipient. Only the intermediary can transfer the ECU given by the donor to the recipient.

ROLE OF THE INTERMEDIARY

The intermediary receives an initial endowment of 80 ECU.

The intermediary's task is to transfer the entirety of the ECU given by the donor to the recipient.

- If the donor has given 25 ECU:

  In situation A, the intermediary can decide either:
  - to transfer the entirety of the 25 ECU to the recipient
  - or to transfer 10 ECU to the recipient and keep 15 ECU for himself/herself

  In situation B, the intermediary can decide either:
  - to transfer the entirety of the 25 ECU to the recipient
  - or to transfer 5 ECU to the recipient and keep 20 ECU for himself/herself

- If the donor has given 0 ECU: The intermediary does not make any decision.

Regardless of the situation, his/her payoff is equal to: 80 ECU + the ECU kept for himself/herself.

Important: For every ECU transferred to the recipient by the intermediary, the recipient receives 2 ECU. For example, if the intermediary transfers 25 ECU, the recipient receives 50 ECU; if the intermediary transfers 5 ECU, the recipient receives 10 ECU.

ROLE OF THE RECIPIENT

The recipient receives an initial endowment of 10 ECU.

The recipient does not make any decision.

Regardless of the situation, his/her payoff is equal to: 10 ECU + (2 x the number of ECU transferred by the intermediary).

Short description of the stages

The session is composed of four stages:

- Stage 1: All the participants answer some questions.
- Stage 2: The donor makes his/her decisions.
- Stage 3: The intermediary makes his/her decisions.
- Stage 4: All the participants answer some questions.

At the end of the session:

- All the participants are informed of the randomly selected situation, of the decisions made by the group members in the randomly selected situation, and of their personal earnings.
- All the participants have to complete a final questionnaire.

Personal Login

When I have finished reading these instructions, please enter your personal login on your screen. It corresponds to the personal login you created yourself when you completed the online questionnaire. As a reminder: we advised you to use "Your mother's or father's first name – his/her day of birth – his/her month of birth" without space or dash. If your mother is called Brigitte and she was born on a 19th of May, it yields "Brigitte1905". Once you have entered your personal login, click "Continue".

Comprehension Questionnaire

You have to complete a comprehension questionnaire. If you have any questions, please raise your hand or press the red button. We will come answer to your questions in private.

***

Once all participants have completed the comprehension questionnaire, the session will start. The role that has been randomly assigned to you will be displayed on your screen. You will then receive more detailed instructions.

[The next set of instructions was distributed after the comprehension questionnaire.]

STAGE 1

In this stage, all the participants have to answer to some questions.

If you are an intermediary or a recipient: You will have to answer the following question: "Among 3 donors randomly selected in today's session, in your opinion how many of these donors will give 25 ECU to the recipient?". You have to enter a number between 0 and 3, inclusive.

You have to answer this question twice: once in situation A, and once in situation B.

If you are a donor or a recipient: You will have to answer to the following question: "Among 3 intermediaries randomly selected in today's session, if their donor decides to give 25 ECU to the recipient, in your opinion how many of these intermediaries will transfer the 25 ECU to the recipient?".

You have to answer to this question twice: once in situation A, and once in situation B.

In total,

1. If you are a donor, you have to answer two questions about the intermediaries' decisions (in situation A and in situation B);
2. If you are an intermediary, you have to answer two questions about the donors' decisions (in situation A and in situation B);
3. If you are a recipient, you have to answer two questions about the donors' decisions (in situation A and in situation B) and two questions about the intermediaries' decisions (in situation A and in situation B).

How do the answers affect your earnings?

At the end of the session, for each role, one of the questions to which you have answered will be randomly selected. If your answer to that question corresponds to what truly happened, you will earn 1€.

Example: Suppose you are a recipient and the question randomly selected is "In situation B, among 3 donors randomly selected in today's session, in your opinion, how many of these donors will choose to give 25 ECU toward the recipient?". The program randomly select 3 donors among the participants to todays' session. If in situation B, x donor(s) among the 3 randomly selected ones, has/have given 25 ECU toward the recipient, then you answer is correct if you answered "x".

STAGE 2

In this stage, the donors make their decisions.

If you are an intermediary or a recipient, you do not make any decision in this stage.

If you are a donor, your task is to decide whether to give 25 ECU or 0 ECU to the recipient.

In total, you have to make two decisions: one in situation A, and one in situation B. However, only one decision will count to determine the payoff of the group members.

Important: When you make your decisions, you do not know which one of your decision will count. You should give the same weight to each of these decisions since you do not know which one will determine the payoffs of the group members.

Which of the donor's decisions determine the payoffs of the group members?

At the end of the session, the computer program will randomly select situation A or situation B. The donor's decision that will count is the decision that was made in the selected situation.

How does the donor's decision affect the payoffs of the group members?

If the donor has chosen to give 0 ECU to the recipient in the randomly selected situation, the payoff of each group member is the following:

- The donor's payoff is 150 ECU.
- The intermediary's payoff is 80 ECU.
- The recipient's payoff is 10 ECU.

If the donor has chosen to give 25 ECU to the recipient in the randomly selected situation:

- The donor's payoff is 125 ECU.
- - The intermediary's and the recipient's payoffs depend on the intermediary's decisions in the third stage.

At the end of the session, you will be informed of the donor's decision in the randomly selected situation.

***

If you have any question, please raise your hand or press the red button. We will come answer to your questions in private.

[The next set of instructions was distributed after stages 1 and 2.]

STAGE 3

In this stage, the intermediaries make their decisions.

If you are a donor or a recipient, you do not make any decision in this stage.

If you are an intermediary, your task is to transfer the entirety of the ECU given by the donor to the recipient.

You have to make several decisions. Look at the screenshot below. There are two pieces of information in bold characters on the screen: these are the two pieces of information that change for each of the decisions.

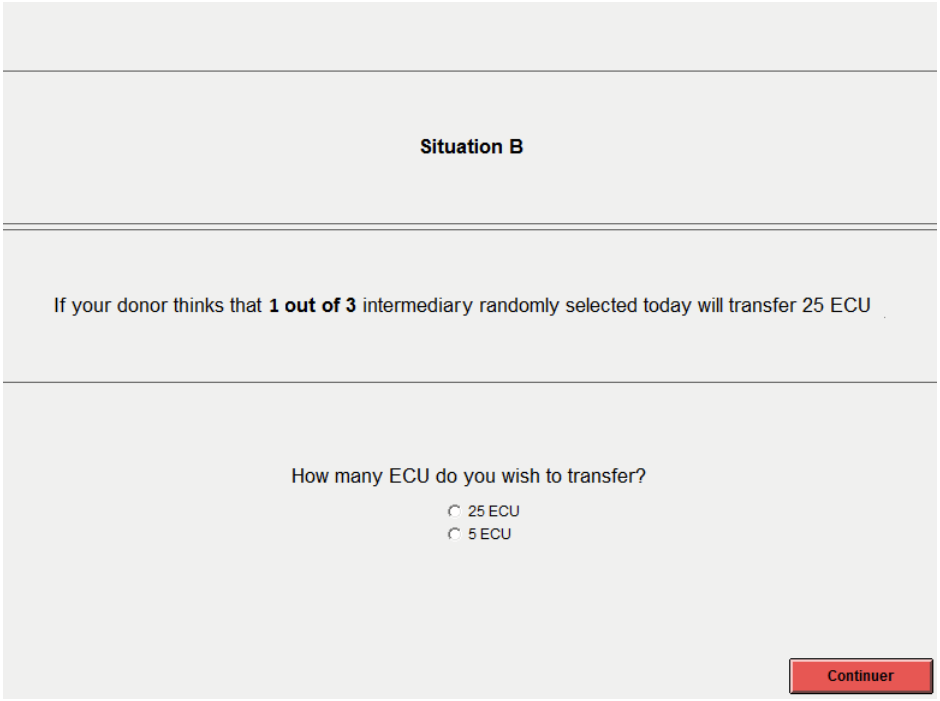


Figure 1.C.1: Screenshot for the *Donor Treatment*

- **Information on the situation**

You have to make a decision in both situation A and situation B. The order of appearance of these situations on your screen is random.

- **Information on your *donor*/*recipient*'s guess**

Remember that in the first stage the *donor*/*recipient* in your group has answered to the following question: "Among 3 intermediaries randomly selected in today's session, if their donor decides to give 25 ECU to the recipient, in your opinion how many of these intermediaries will transfer the 25 ECU to the recipient?". There were four possible answers: 0, 1, 2 or 3. You have to make a decision for each of the possible answers.

When you make your decisions, you do not know how many ECU the donor in your group has decided to give to the recipient. You have to make your decisions assuming that the donor has given 25 ECU.

In total, you have to make eight decisions: four decisions corresponding to the four possible answers of the *donor*/*recipient* in your group in situation A, and four decisions cor-



Figure 1.C.2: Screenshot for the *Recipient Treatment*

responding to the four possible answers of the *donor*/*recipient* in your group in situation B.

Important: When you make your decisions, you do not know which one of your decision will count. You should give the same weight to each of your decisions since you do not know which one will determine the payoff of the group members.

Which of the intermediary's decisions will determine the payoff of the group members?

- If the donor has chosen to give 0 ECU to the recipient: none of the intermediary's decisions will determine the payoff of the group members.
- If the donor has chosen to give 25 ECU to the recipient: one of the intermediary's decisions will determine the payoff of the group members.

At the end of the session, the computer program will randomly select situation A or situation B. Among the intermediary's decisions made in the randomly selected situation, the computer program selects the decision corresponding to the answer given by the *donor*/*recipient* of your group in the first stage. It is this decision that determines the payoff of the group

members.

Example: Suppose that the program randomly selects situation A. Suppose then that, to the question "In situation B, among 3 intermediaries randomly selected in today's session, if their donor decides to give 25 ECU to the recipient, in your opinion how many of these intermediaries will transfer 25 ECU to the recipient?", the *donor*/*recipient* of your group has answered "x". Then, the program selects the decision made by the intermediary when his/her screen displayed "Situation B" and "Your *donor*/*recipient* believes that x intermediaries among 3 randomly selected today will transfer 25 ECU."

How does the intermediary's decision affect the payoff of the group members?

If the donor has given 25 ECU to the recipient in the randomly selected situation, one of the intermediary's decisions determines the payoffs of the group members.

The intermediary may have made three types of decisions:

- Regardless of the situation, if the intermediary transfers 25 ECU to the recipient, the intermediary's payoff is 80 ECU and the recipient's payoff is 60 ECU.
- If situation A is randomly selected and if the intermediary transfers 10 ECU to the recipient and keeps 15 ECU for himself/herself, the intermediary's payoff is 95 ECU and the recipient's payoff is 30 ECU.
- If situation B is randomly selected and if the intermediary transfers 5 ECU to the recipient and keeps 20 ECU for himself/herself, the intermediary's payoff is 100 ECU and the recipient's payoff is 20 ECU.

At the end of the session, you will be informed of the donor's decision in the randomly selected situation.

***

If you have any questions, please raise your hand or press the red button. We will come answer to your questions in private.

[The next set of instructions was distributed after the stage 3]

STAGE 4

1) First, all the participants have to answer to questions of type 1.

You have to evaluate the different possible decisions of a donor and of an intermediary. More precisely, for each possible decision of a donor or of an intermediary, you are asked to indicate whether this decision is socially appropriate and consistent with moral or proper social behavior, or socially inappropriate and inconsistent with moral or proper behavior.

Consider that a decision is socially appropriate if the majority of people agree to say that it is the correct or ethical thing to do. You have to rate each decision using the following scale: very socially inappropriate, somewhat socially inappropriate, somewhat socially appropriate or very socially appropriate.

2) Then, the donor and the intermediary have to answer to questions of type 2.

You are asked to guess the decision made by a participant earlier in the session.

How do the answers affect your earnings?

At the end of the session, for each role, the program will randomly select one of the questions to which you have answered in this stage. If you are a recipient, the randomly selected question is for sure a question of type 1. If you are a donor or an intermediary, the question randomly selected can be a question of type 1 or a question of type 2.

- If the randomly selected question is a question of type 1:

Your earning depends on the answers of the other participants in the same role as you in today's session. The computer program determines the answer given by the highest number of participants in the same role as you (you included) to this question. You earn €1 if your answer corresponds to the answer the most frequently given by participants in the same role as you. In case of a tie between two answers, the program randomly selects one of the tie answers.

Example: Suppose there are six participants in today's session who have the role of donors. A question of type 1 is randomly selected. To that question, one donor has answered " very socially inappropriate", two donors have answered "somewhat socially appropriate" and three donors have answered "very socially appropriate". The answer the most frequently given by the donors is "very socially appropriate". Then, the three donors who have answered "very socially appropriate" earn €1, the other donors earn nothing.

- If the randomly selected question is a question of type 1:

If you have guessed correctly a previous decision, you earn €1.

END OF THE SESSION

At the end of the session, you will be informed of the situation randomly selected, of the decisions made by your group members in the randomly selected situation, and of your personal payoff. Then, you will be asked to complete a final questionnaire.

At the end of the session, please remain seated and silent until an experimenter invites you to proceed to the payment room. At this moment, bring only your computer tag and your payment receipt completed with you.

***

If you have any questions, please raise your hand or press the red button. We will come answer your questions in private.

1.C.2 Instructions for the Online Questionnaire [Translated from French]

PART 0 - Introduction

Thank you for accepting to answer this questionnaire in order to complete your registration to the experiment. Answering to this questionnaire will take approximately 10 minutes. Please read carefully each sentence and remain concentrated. We are interested in your genuine answers, not what you think you should answer.

PART 1 - GASP Questionnaire (Cohen et al., 2011)

Here are situations that people are likely to encounter in day-to-day life, followed by common reactions to those situations. As you read each scenario, try to imagine yourself in that situation.

Please indicate the likelihood that you would react in the way described by using the following categories: (1) Very Unlikely, (2) Unlikely, (3) Slightly Likely, (4) Unlikely, (5) About 50% Likely, (6) Slightly Likely, (7) Very Likely.

1. After realizing you have received too much change at a store, you decide to keep it because the salesclerk does not notice. What is the likelihood that you would feel uncomfortable about keeping the money?
2. You are privately informed that you are the only one in your group that did not make the honor society because you skipped too many days of school. What is the likelihood that this would lead you to become more responsible about attending school?
3. You rip an article out of a journal in the library and take it with you. Your teacher discovers what you did and tells the librarian and your entire class. What is the likelihood that this would make you would feel like a bad person?
4. After making a big mistake on an important project at work in which people were depending on you, your boss criticizes you in front of your co-workers. What is the likelihood that you would feign sickness and leave work?
5. You reveal a friend's secret, though your friend never finds out. What is the likelihood that your failure to keep the secret would lead you to exert extra effort to keep secrets in the future?
6. You give a bad presentation at work. Afterwards your boss tells your co-workers it was your fault that your company lost the contract. What is the likelihood that you would feel incompetent?
7. A friend tells you that you boast a great deal. What is the likelihood that you would stop spending time with that friend?

8. Your home is very messy and unexpected guests knock on your door and invite themselves in. What is the likelihood that you would avoid the guests until they leave?

9. You secretly commit a felony. What is the likelihood that you would feel remorse about breaking the law?

10. You successfully exaggerate your damages in a lawsuit. Months later, your lies are discovered and you are charged with perjury. What is the likelihood that you would think you are a despicable human being?

11. You strongly defend a point of view in a discussion, and though nobody was aware of it, you realize that you were wrong. What is the likelihood that this would make you think more carefully before you speak?

12. You take office supplies home for personal use and are caught by your boss. What is the likelihood that this would lead you to quit your job?

13. You make a mistake at work and find out a co-worker is blamed for the error. Later, your co-worker confronts you about your mistake. What is the likelihood that you would feel like a coward?

14. At a co-worker's housewarming party, you spill red wine on their new cream-colored carpet. You cover the stain with a chair so that nobody notices your mess. What is the likelihood that you would feel that the way you acted was pathetic?

15. While discussing a heated subject with friends, you suddenly realize you are shouting though nobody seems to notice. What is the likelihood that you would try to act more considerately toward your friends?

16. You lie to people but they never find out about it. What is the likelihood that you would feel terrible about the lies you told?

Guilt Negative-Behavior-Evaluation (NBE)	1, 9, 14, 16
Guilt Repair (R)	2, 5, 11, 15
Shame Negative-Self-Evaluation (NSE)	3, 6, 10, 13
Shame Withdraw (W)	4, 7, 8, 12

Table 1.C.1: GASP Questionnaire - Answers Key

PART 2 - Honesty-Humility Scale from the 100-items HEXACO Personality Inventory - Revised (Lee and Ashton, 2004)

Please indicate how much you agree or disagree with these statements about you by using the following categories: (1) Strongly disagree, (2) Disagree, (3) Neutral (neither agree nor disagree), (4) Agree, (5) Strongly disagree.

1. If I want something from a person I dislike, I will act very nicely toward that person in order to get it.
2. If I knew that I could never get caught, I would be willing to steal a million dollars.
3. Having a lot of money is not especially important to me.
4. I am an ordinary person who is no better than others are.
5. I would not use flattery to get a raise or promotion at work, even if I thought it would succeed.
6. I would be tempted to buy stolen property if I were financially tight.
7. I would like to live in a very expensive, high-class neighborhood.
8. I would not want people to treat me as though I were superior to them.
9. If I want something from someone, I will laugh at that person's worst jokes.
10. I would never accept a bribe, even if it were very large.
11. I would like to be seen driving around in a very expensive car.
12. I think that I am entitled to more respect than the average person is.
13. I would not pretend to like someone just to get that person to do favors for me.
14. I would be tempted to use counterfeit money, if I were sure I could get away with it.
15. I would get a lot of pleasure from owning expensive luxury goods.
16. I want people to know that I am an important person of high status.

Sincerity	1R, 5, 9R, 13
Fairness	2R, 6R, 10, 14R
Greed-Avoidance	3, 7R, 11R, 15R
Modesty	4, 8, 12R, 16R

Table 1.C.2: Honesty-Humility Scale - Answers Key[25]

PART 3 – Inspired by the Self Report Altruism Scale (Rushton et al., 1981)[26]

Please indicate the frequency with which you have carried out the following acts by using the following categories: (1) Never, (2) Once, (3) More than once, (4) Often, (5) Very Often.

[26]Three items were excluded: "I have made change for a stranger", "I have given a stranger a lift in my car" and "I have bought 'charity" Christmas cards deliberately because I knew it was a good cause".

1. I have helped a stranger change a flat tire.[27]
2. I have given directions to a stranger.
3. I have given money, goods or clothes to a charity.[28]
4. I have delayed an elevator and held the door open for a stranger.
5. I have donated blood.
6. I have helped carry a stranger's belongings (books, parcels, etc.).
7. I have allowed someone to go ahead of me in a lineup (at photocopy machine, in the supermarket).
8. I have pointed out a clerk's error (in a bank, at the supermarket) in undercharging me for an item.
9. I have let a neighbor whom I did not know too well borrow an item of some value to me (e.g., a dish, tools, etc.)
10. I have done volunteer work for a charity.
11. I have helped a classmate who I did not know that well with a homework assignment when my knowledge was greater than his or hers.
12. I have before being asked, voluntarily looked after a neighbor's pets or children without being paid for it.
13. I have offered to help a handicapped or elderly stranger across a street.
14. I have offered my seat on a bus or train to a stranger who was standing.
15. I have helped an acquaintance to move households.
16. I have given money to a stranger who needed it (or asked me for it).

PART 4 – Socio-Demographics

1. Risk Preferences (Dohmen et al., 2011)
   How would you describe yourself? Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks? Please tick a box on the scale, where the value "0" means "not at all willing to take risks" and the value "10" means "very willing to take risks".

[27] Originally: "I have helped push a stranger's car out of the snow."

[28] Originally it was two different items: « I have given money to charity" and "I have donated goods or clothes to a charity".

2. Time Preferences (Visher et al., 2013)
   How would you describe yourself? Are you generally an impatient person, or someone who always shows great patience? Please tick a box on the scale, where the value "0" means "very impatient" and the value "10" means "very patient".

3. Religiosity
   How would you describe yourself? How often do you pray?
   - I never pray
   - I seldom pray
   - I pray every week
   - I pray more than once a day

4. Gender
   Please indicate your gender.
   - Female
   - Male

5. Age
   Please indicate your age.

6. Status
   Please indicate your status.
   - Student
   - Employed
   - Unemployed
   - Retired

   (a) School - *if your answer to question 6 is "Student"*
   Which school do you attend?
   - EM Lyon
   - Ecole Centrale Lyon
   - ISOstéo
   - Université Lyon 1
   - Université Lyon 2
   - Université Lyon 3
   - Université Catholique de Lyon
   - Other

   (b) Field of Study - *if your answer to question 6 is "Student"*
   What is your field of study?

- Economics and Management
- Social Sciences
- Arts and Humanities
- Engineering Sciences
- Medical Studies
- Other

(c) Professional Activity - *if your answer to question 6 is "Employed"*
What is your current professional status?

- Farmer
- Craftsman, shopkeeper, business owner
- Executive and higher intellectual occupations
- Civil servant, administrative employee
- Employee
- Worker

7. Number of previous experiments
In how many GATE-LAB experimental sessions have you participated already?

8. Personal Login
Please choose a personal login. Choose a login that you can remember easily since you will need this login to start the experimental session. We suggest you use "Mother's or Father's first name - her/his day of birth - her/his month of birth" without space or dash. For example, if your mother is called Brigitte and is born a May 19th, the suggested login is "Brigitte1905".

Bibliography

Balafoutas, L. and Fornwagner, H. (2017). The limits of guilt. *Journal of the Economic Science Association*, 3(2):137–148.

Bellemare, C., Sebald, A., and Strobel, M. (2011). Measuring the willingness to pay to avoid guilt: estimation using equilibrium and stated belief models. *Journal of Applied Econometrics*, 26(3):437–453.

Bellemare, C., Sebald, A., and Suetens, S. (2018). Heterogeneous guilt sensitivities and incentive effects. *Experimental Economics*, 21(2):316–336.

Bracht, J. and Regner, T. (2013). Moral emotions and partnership. *Journal of Economic Psychology*, 39:313–326.

Khalmetski, K., Ockenfels, A., and Werner, P. (2015). Surprising gifts: Theory and laboratory evidence. *Journal of Economic Theory*, 159:163–208.

Moulton, R. W., Burnstein, E., Liberty Jr, P. G., and Altucher, N. (1966). Patterning of parental affection and disciplinary dominance as a determinant of guilt and sex typing. *Journal of Personality and Social Psychology*, 4(4):356.

Patel, A. and Smith, A. (2019). Guilt and participation. *Journal of Economic Behavior & Organization*, 167:279–295.

Peeters, R. and Vorsatz, M. (2018). Simple guilt and cooperation. *University of Otago Economics Discussion Papers*, 1801.

Regner, T. and Harth, N. S. (2014). Testing belief-dependent models. *Jena Economic Research Papers*.

Chapter 2

Guilt Aversion and Vulnerability

This chapter is co-authored with Giuseppe Attanasi and Marie Claire Villeval.

2.1 Introduction

Based on psychological insights (Baumeister et al., 1994), economists have modelled how guilt can influence actions. Within the framework of psychological game theory[1], Battigalli and Dufwenberg (2007) define guilt aversion as a belief-dependent motivation: an agent suffers a psychological cost, *i.e.*, feels guilty, if he lets down others' expectations. Correspondingly, a plethora of psy-game theory-driven experiments have focused on guilt aversion as a potential driver of pro-social behavior in social dilemma games (see the survey of Battigalli and Dufwenberg, 2020). The overwhelming majority of these experiments are based on two social dilemma games, the dictator and the trust games.[2]

[1]This theory departs from traditional game theory in assuming that players' utilities do not only depend on their decisions but also on their beliefs about decisions, beliefs, or information (Geanakoplos et al., 1989; Battigalli and Dufwenberg, 2009).

[2]Considered together, the dictator game and the trust game currently represent, to the best of our knowledge, the focus of 75% of published psy-game experimental studies on guilt aversion (see Table 2.A.1 in Appendix).

A common feature of these two games, as used in this literature, is that co-players are vulnerable, that either their final payoff or the use of their initial endowment depends on the actions of the decision-maker. However, both the necessity of having vulnerable co-players to induce guilt aversion and the potential influence of the nature of this vulnerability, have never been experimentally tested. The present study addresses these two questions.

Indeed, guilt aversion has been shown to be modulated by a series of factor. It is influenced by the communication of others' expectations as well as by the very nature of these expectations. The possibility for players to communicate greatly facilitates the expression of the trustee's guilt aversion, as evidenced in the milestone paper of Charness and Dufwenberg (2006) and replicated in many experimental papers since (*e.g.*, Attanasi et al., 2013; Bracht and Regner, 2013, Kawagoe and Narita, 2014; Balafoutas and Sutter, 2017; Attanasi et al., 2019a). Turning to the nature of expectations, "reasonable" expectations appear more likely to be taken into account by guilt-averse players. Balafoutas and Fornwagner (2017), Khalmetski (2016) and Danilov et al. (2019) all reported an inverse-U shaped relationship between second-order beliefs and sharing decisions: dictators are less pro-social when they deem that recipients expect to receive too little or too much. Further, the emergence of trustees' guilt aversion is facilitated by the perceived legitimacy of trustors' normative expectations (Andrighetto et al., 2015; Pelligra et al., 2020).

As for the role of vulnerability in the return decisions of trustees, recent studies by Cox et al. (2016) and Engler et al. (2018) showed that trustees' returns increase with the vulnerability of the trustor.[3] However, with regard to guilt aversion, the primitive of

[3] Cox et al. (2016) considered that the trustor is vulnerable if she made a choice such that the maximum payoff she can obtain—assuming that the trustee is selfish—is lower than the maximum payoff she could

Battigalli and Dufwenberg (2007) model is that the dictator (resp., the trustee) can feel guilty toward a vulnerable recipient (resp., trustor). Hence, by assuming the vulnerability of the co-player, they (and the following applications of their model) never questioned the influence of vulnerability on the emergence of guilt aversion.

A thorough examination of the literature suggests, however, that it may be important to raise this issue. It is noteworthy that most of the empirical evidence in favor of guilt aversion is based on two games that rely on different natures of co-players' vulnerability. In the dictator game, the vulnerability of the recipient can be characterised as ex-post. We define a player as *ex-post vulnerable* if her material payoff depends on the actions of the decision-maker. In contrast, in the trust game, the vulnerability of the trustor is both ex-post and ex-ante. We define a player as *ex-ante vulnerable* if her initial endowment can be entrusted to the decision-maker. Bellemare et al. (2017) contrasted the two games in a single study and found no difference in the intensity of guilt aversion. This suggests that the combined effect of both types of vulnerability (in the trust game) is not additive, although we lack a comparison with only ex-ante vulnerable co-players. A first step in this direction has been taken by Attanasi et al. (2019b) (Chapter 1) who compared guilt aversion toward ex-ante vulnerable co-players *vs.* ex-post vulnerable co-players. They reported no difference be it in the proportion of guilt averse players or in the intensity of the observed guilt aversion. Their results provide indirect evidence that none of the two types of vulnerability of co-player is a necessary condition to trigger guilt. Yet, altogether no final conclusion can be drawn from these studies as they all lack a control condition with no vulnerability at all and they do not allow a comparison, in a single

have obtained otherwise—again assuming a selfish trustee. Engler et al. (2018) defined three degrees of vulnerability in a trust game: the trustor is either (i) not vulnerable if she made a choice such that the minimum payoff she can obtain by entrusting her endowment is higher than the payoff she could have obtained by not entrusting it; (ii) vulnerable if she made a choice such that the two payoffs she can obtain by entrusting her endowment are respectively lower and higher than the payoff she could have obtained by not entrusting it; (iii) very vulnerable if she made a choice such that the maximum payoff she can obtain by entrusting her endowment is lower than the payoff she could have obtained not entrusting it.

study, of all the possible combinations of ex-ante and ex-post vulnerability.

Having this in mind, in the present study, we build on Attanasi et al. (2019b) (Chapter 1) and introduce four variations of a Trust mini-game with a passive player (Quasi-Trust mini-games, henceforth) that allow us to systematically compare the four possible combinations of vulnerability: no vulnerability, ex-ante vulnerability, ex-post vulnerability, ex-ante and ex-post vulnerabilities. Secondarily, this design offers the possibility to test whether observing or not the intentions of a vulnerable player makes a difference in the willingness to avoid to disappoint her (by comparing an active player (A) whose intentions are observable and a passive player (C) with the same type of vulnerability).

The four Quasi-Trust mini-games are: the Investment game, the Reversed-Investment game, the Donation game (similar to Attanasi et al., 2019b), and the Exploitation game. In each game the second mover (B) can be entrusted by the first mover (A) with a sum of money coming from the endowment of another player (A or C, depending on the game); then, he can redistribute this money between himself and another player (A or C, depending on the game).[4] The four Quasi-Trust mini-games are highly comparable since they share: for each player, the same initial endowment; for each of the two active players, the same set of strategies; for the potentially guilt averse player B, the same material payoff given the game terminal node (and so the same best-reply function, if he is selfish). These games differ only in which player's vulnerability and which type of vulnerability is activated: ex-ante and ex-post vulnerabilities can be activated for the same player (A or C) – leaving the second one not vulnerable at all – or can be distributed between players A and C (A: ex-ante and C: ex-post or the reverse).

[4] In each game, players A and C are denoted as female ("she") and player B denoted as male ("he").

Decisions of player B, which are the focus of the present study, can then be contrasted across the four games (*i.e.* across the four combinations of vulnerability), which are played within-participants. Between-participants, player B's decisions are also elicited either conditional on the first-order beliefs of player A (active) or of player C (passive). Therefore, we have a 4x2 design, which allows to test the (in)dependence of guilt aversion from the co-players' vulnerability and status.

From a theoretical point of view, we rely on a portable model of lexicographic altruism and role-dependent guilt which provides predictions for the entire set of games. We assume that both players A and B can be altruistic toward the most disadvantaged player, while only player B can feel guilty. As for the latter, we assume no influence of the partners' vulnerability on triggering guilt: player B may feel guilty even if the partner is not vulnerable and even if he cannot observe the partner's intentions, such as when the latter is a simple observer. Guilt sensitivity is mainly triggered by the role in the game.

Our experimental results show no significant difference in the proportion of guilt-averse B-subjects across Quasi-Trust mini-games, with a relevant fraction of B-subjects expressing guilt aversion even toward a player who is not vulnerable. This lack of significant difference suggests that vulnerability and its nature do not modulate the trustee's guilt aversion in a Quasi-Trust game. We interpret such insensitivity of guilt aversion to the co-player's vulnerability as further support to guilt mainly being role-dependent in two-stage games with asymmetric roles (as suggested by Attanasi et al., 2016).

The remainder of the paper is organized as follows. Section 2.2 presents the rationale for our four (new) games given our empirical interest in the impact of the partners' vulnerability. Section 2.3 introduces our theoretical model and related predictions.

Section 3.3 describes the experimental design. Section 2.5 presents the experimental results and Section 2.6 concludes.

2.2 The Quasi-Trust Mini-Games

To manipulate vulnerability, we introduce four Quasi-Trust games with three players: the Investment game (Figure 2.2.1), the Reversed-Investment game (Figure 2.2.2), the Donation game (Figure 2.2.3) and the Exploitation game (Figure 2.2.4). In each game, players A and B are active while player C is passive. Players' material payoffs in Figures 2.2.1-2.2.4 are shown according to the players' alphabetical order.

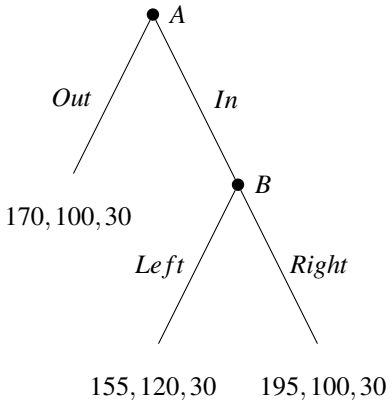


Figure 2.2.1: The Investment Game

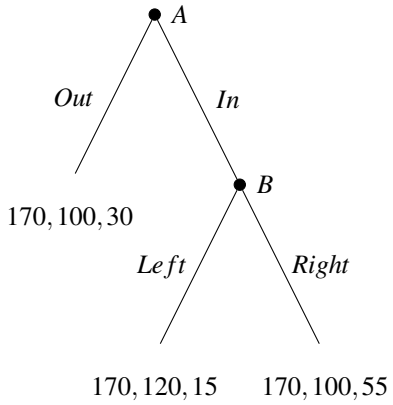


Figure 2.2.2: The Reversed-Investment Game

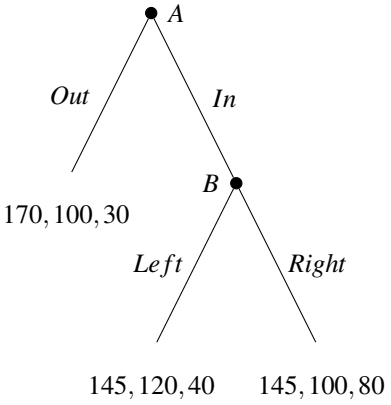


Figure 2.2.3: The Donation Game

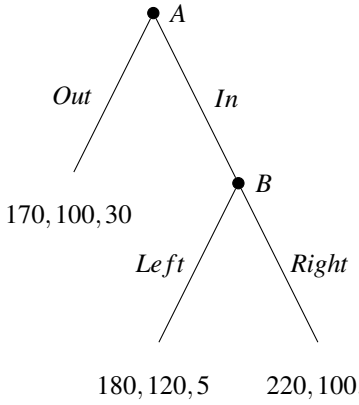


Figure 2.2.4: The Exploitation Game

Each game unfolds as follows. A is the first mover, she can choose *In* or *Out*. If A chooses *Out*, the game ends with material payoffs corresponding to the players' initial endowment (170 ECU for A, 100 ECU for B, 30 ECU for C).[5] If A chooses *In*, she sends 25 ECU to B, with this amount being taken either from player A's endowment or player C's endowment (ex-ante vulnerability), depending on the game. After *In*, player B decides how to allocate the 25 ECU between himself and another player, this player being A or C (ex-post vulnerability), depending on the game. In particular, if B chooses *Left*, he transfers 5 ECU to another player and keeps 20 ECU for himself; if B chooses *Right*, he transfers the 25 ECU to this other player. Each ECU transferred by B to another player (A or C, depending on the game) is doubled, which captures the positive externality of trust.[6]

[5] All material payoffs are expressed in Experimental Currency Units (ECU) where 10 ECU = €1 (see the experimental procedures in Section 2.4.3).

[6] Several game-independent features of the final distributions of material payoffs are worth noting. First, given the terminal node, B's material payoff is the same across the four games: if B chooses *Right* after *In*, his material payoff corresponds to his initial endowment (*Out*); if B chooses *Left* after *In*, his material payoff corresponds to his initial endowment plus the 20 ECU that he takes for himself. However, the payoff manipulation across the four games affects A's and C's payoffs (see Figure 2.2.1 to Figure 2.2.4). Next, no decision can lead to the equalization of payoffs between two or three players. Hence, no payoff distribution should be more salient than others. Furthermore, the ranking of payoffs cannot be affected by the players' decisions, which limits social comparison motives in decision making. Finally, the total surplus at a given terminal node is the same across games, this way keeping efficiency concerns constant across games.

In the **Investment game** (Figure 2.2.1), A can entrust B with 25 ECU taken from her own endowment. B then decides how to allocate these 25 ECU between A and himself. In this game, B's choice affects both the use of A's initial endowment and A's material payoff but it does not concern C, *i.e.*, A is both ex-ante and ex-post vulnerable whereas C is not vulnerable. The Investment game is a simplified version (mini-game) of the classical Trust game (see Berg et al., 1995; Buskens and Raub, 2013; Attanasi et al., 2016), with the additional feature of an external observer, C, whose payoff is not affected by A (trustor) and B's (trustee) actions.

In the **Reversed-Investment game** (Figure 2.2.2), A can entrust B with 25 ECU taken from C's endowment. B then decides how to allocate these 25 ECU between C and himself. In this game, B's choice affects both the use of C's initial endowment and C's material payoff but it does not concern A, *i.e.*, A is not vulnerable and C is both ex-ante and ex-post vulnerable. Thus, the Reversed-Investment game is a modified version of the Investment game where all monetary consequences of A's investment choice fall on C's payoff: A invests C's endowment and the doubled amount can enrich C.

In the **Donation game** (Figure 2.2.3), A can entrust B with 25 ECU taken from her own endowment. B then decides how to allocate these 25 ECU between C and himself. In this game, B's choice affects the use of A's initial endowment as well as C's material payoff, *i.e.*, A is ex-ante vulnerable and C is ex-post vulnerable. Thus, the Donation game is a modified version of the Investment game where the positive monetary consequences of A's investment choice fall on C's payoff: A invests her endowment and the doubled amount can enrich C. This is similar to the Embezzlement game of Attanasi et al. (2019b).

In the **Exploitation game** (Figure 2.2.4), A can entrust B with 25 ECU taken from C's endowment. B then decides how to allocate these 25 ECU between A and himself. In this game, B's choice affects the use of C's initial endowment as well as A's material payoff, *i.e.*, A is ex-post vulnerable and C is ex-ante vulnerable. Thus, the Exploitation game is a modified version of the Investment game where the negative monetary consequences of A's investment choice fall on C's payoff: A invests C's endowment and the doubled amount can enrich A.

Figure 2.2.5 summarizes the manipulation of A's and C's vulnerability across the four games.

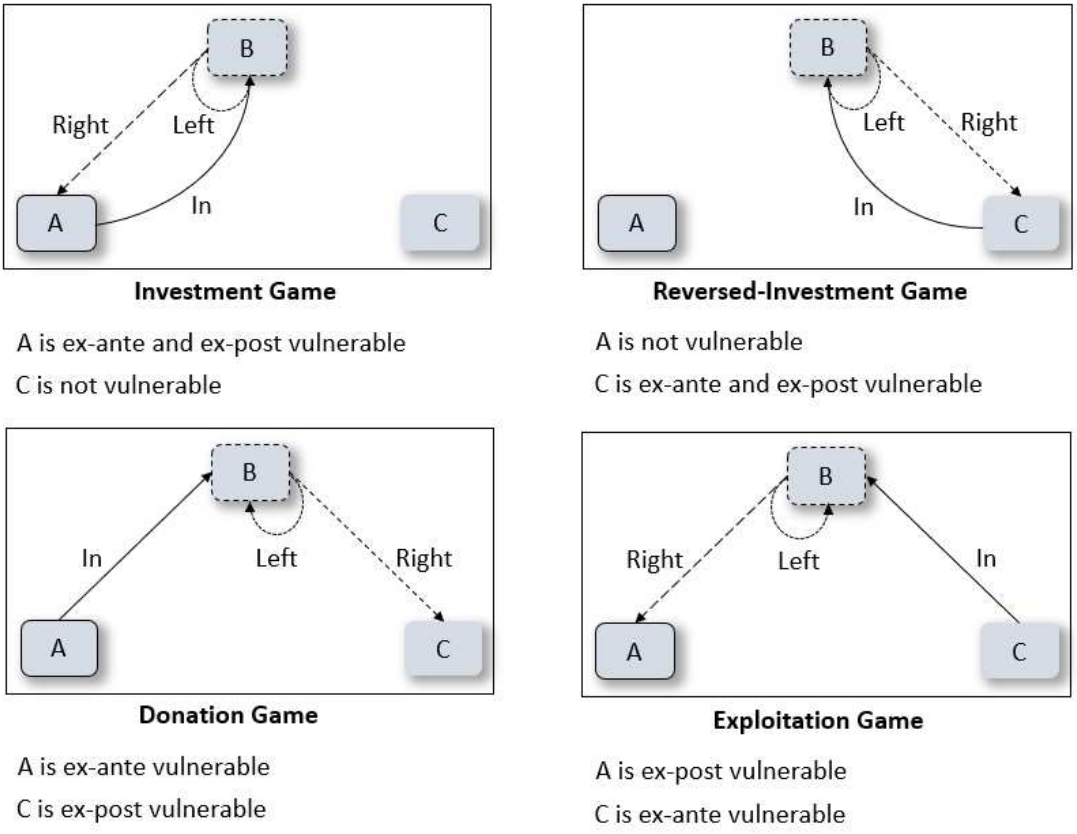


Figure 2.2.5: Vulnerability in the four Quasi-Trust mini-games

Notes: In each panel, plain lines indicate which player's endowment is used by A to transfer money to B through strategy *In*; dashed lines indicate player B's strategies. Short dashes indicate strategy *Left*, with only 5 out of 25 ECU transferred to another player, and the rest kept by B. Long dashes indicate strategy *Right*, with all the 25 ECU transferred to another player, thus generating higher positive externalities, since each ECU transferred by B is doubled.

2.3 Theoretical Model and Hypotheses

In this section, we develop a theoretical model of lexicographic altruism and role-dependent guilt based on the work of Attanasi et al. (2019b). After describing the players' utility functions, we analyze A's and B's best-reply functions. Finally, we elaborate theory-driven experimental hypotheses on A's and B's behavior. We denote player j's material payoff as π_j, with $j \in \{A, B, C\}$, at each terminal node $z \in \{O, L, R\}$ of

the games, *i.e.*, respectively, for each terminal history *Out*, *Left* after *In*, and *Right* after *In*.

2.3.1 Utility Functions

Since C is passive, we assume that she is purely self-interested. Therefore, **C's utility function** coincides with her material payoff, *i.e.*, $U_C(z) = \pi_C(z)$ for each $z \in \{O, L, R\}$. This assumption is motivated by the fact that, in each game and for each terminal history, C always gets the lowest payoff in the triplet.

As for **A's utility function**, we assume that she can be altruistic toward both B and C, since, at each terminal node z, A always gets the highest payoff independently from the game and the strategy profile in that game. We also assume that A's altruistic preferences toward disadvantaged players are lexicographic. Precisely, since C is always the most disadvantaged player and B is always the second most disadvantaged player, A is altruistic only toward C when C's payoff depends on A's strategy, and only toward B when B's payoff depends on A's strategy but C's payoff does not.[7] Therefore, A can be altruistic toward player C in the Reversed-Investment, Donation and Exploitation games, and she can be altruistic toward B in the Investment game.

We model A's feeling of altruism toward player $h \in \{B, C\}$, F_{Ah}, as A's utility derived from the payoff of h. It is the product of two terms: $\phi_{Ah} \geq 0$, A's sensitivity to altruism toward h, and $\pi_h(z)$, h's material payoff. With this, A's utility (Eq. (2.1)) is composed of

[7] The assumption of lexicographic altruistic preferences is broadly consistent with inequity-aversion models (Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999), since C is the most disadvantaged player.

her material payoff and her feeling of altruism toward $h \in \{B,C\}$ (Eq. (2.2)):

$$U_A(\phi_{Ah},z) = \pi_A(z) + F_{Ah}(\phi_{Ah},z) \tag{2.1}$$

$$\text{where } F_{Ah}(\phi_{Ah},z) = \phi_{Ah} \cdot \pi_h(z) \tag{2.2}$$

with $h = B$ in the Investment game and $h = C$ in the remaining three games.

Let us now introduce **B's utility function**. Besides B's concern for his own payoff, π_B, we assume that B has lexicographic altruistic preferences toward disadvantaged players modeled like those of player A (see F_{Bh}, namely B's utility derived from the payoff of player $h \in \{A,C\}$, in Eq. (2.4)). With this, by construction of the four games, since C is always the most disadvantaged player and A is always the most advantaged player, B can be altruistic only toward C: he is altruistic only toward C when C's payoff depends on his strategy, and toward no player when C's payoff does not depend on his strategy (in the latter case, there is no player more disadvantaged than him whose payoff he can increase). Therefore, F_{Bh} in Eq. (2.4) essentially coincides with F_{BC} in each of the four games ($h = C$). The latter only has a strategic impact in the Reversed-Investment and Donation games, where C's material payoff depends on B's strategy: from Eq. (2.4), $F_{BC}(\phi_{BC},R) > F_{BC}(\phi_{BC},L)$, *i.e.*, *Right* after *In* is a more altruistic strategy than *Left* after *In*. In the Investment and Exploitation games, where C's material payoff does not depend on B's strategy, it is $F_{BC}(\phi_{BC},R) = F_{BC}(\phi_{BC},L)$, hence B's altruism is irrelevant.

Furthermore, in line with the role-dependent guilt model of Attanasi et al. (2016), we assume that B can feel guilty due to his role in the game, whereas A does not.[8]

[8]See the discussion in Attanasi et al. (2016), p. 649, where they argue that role dependence of guilt preferences is plausible in asymmetric games (see, *e.g.*, Attanasi et al., 2013, 2019a, for indirect experimental evidence corroborating this assumption). In particular, they discuss how the assumption that sensitivity to guilt is triggered only when playing in the role of trustee (and not in the role of trustor) in the trust game resonates with the evolutionary psychology of emotions and the conceptual act theory of

B's feeling of guilt, G_{Bjk}, with $j,k \in \{A,C\}$ in Eq. (2.5), represents his disutility derived from letting down j's beliefs on the strategy he will select, which will affect k's payoff, with j not necessarily equal to k. More precisely, it is the product of two terms: $\gamma_{Bjk} \geq 0$, B's guilt sensitivity about j's beliefs when B's strategy affects k's payoff; and the difference, if positive, between j's beliefs about k's payoff after *In*, $\mathbb{E}_j[\pi_k(z|In)]$, and k's actual material payoff after *In*, $\pi_k(z|In)$. More precisely, if $\mathbb{E}_j[\pi_k(z|In)] = \alpha_{jB} \cdot \pi_k(R) + (1-\alpha_{jB}) \cdot \pi_k(L) > \pi_k(z|In)$ (where α_{jB} is j's first-order belief that B chooses *Right* after *In*), then B feels guilty from letting down j's beliefs on k's payoff; otherwise his guilt feeling G_{Bjk} is null since he does not let down j's beliefs on k's payoff. With this, B's utility after *In* is represented in Eq. (2.3), with altruism and guilt feelings represented in respectively Eq. (2.4) and Eq. (2.5) for $j,k \in \{A,C\}$:

$$U_B(\phi_{BC}, \gamma_{Bjk}, \alpha_{jB}, z|In) = \pi_B(z|In) + F_{BC}(\phi_{BC}, z|In) - G_{Bjk}(\gamma_{Bjk}, \alpha_{jB}, z|In) \quad (2.3)$$

$$\text{where } F_{BC}(\phi_{BC}, z|In) = \phi_{BC} \cdot \pi_C(z|In) \quad (2.4)$$

$$\text{and } G_{Bjk}(\gamma_{Bjk}, \alpha_{jB}, z|In) = \gamma_{Bjk} \cdot \max\{0, \mathbb{E}_j[\pi_k(z|In)] - \pi_k(z|In)\} \quad (2.5)$$

We anticipate here that we implement an experimental design where the impact of guilt sensitivity toward A can be analyzed separately from the impact of guilt sensitivity toward C (see Section 3.3). In fact, we use a between-subject design to elicit B's belief-dependent strategy conditional on either A's (treatment A) or C's (treatment C) first-order beliefs about *Right* if *In*. Therefore, as for Eq. (2.5), we elicit guilt sensitivity γ_{Bjk} with $j = A$ in treatment A and $j = C$ in treatment C regardless of the Quasi-Trust mini-game. In treatment A, where B's strategy is elicited conditional on A's first-order beliefs ($j = A$, hence G_{BAk}), the standard Battigalli and Dufwenberg (2007) definition of guilt aversion ($k = A$, hence G_{BAA}) only applies in the Investment and Exploitation games, while

emotion. Similar arguments can be provided in support of sensitivity to guilt being triggered only when playing in the role of player B (and not in the role of player A) in our four Quasi-Trust mini-games of Figure 2.2.1 to Figure 2.2.4.

the extended definition ($k = C$, hence G_{BAC}) also applies in the Reversed-Investment and Donation games. Correspondingly, in treatment C, where B's strategy is elicited conditional on C's first-order beliefs ($j = C$, hence G_{BCk}), the standard Battigalli and Dufwenberg (2007) definition of guilt aversion ($k = C$, hence G_{BCC}) only applies to the Reversed-Investment and Donation games, and the extended definition ($k = A$, hence G_{BCA}) also applies to the Investment and Exploitation games.

2.3.2 Best-Reply Analysis

We elaborate our hypotheses relying on best-reply analysis rather than on Bayesian equilibrium. Indeed, a standard equilibrium analysis has no compelling foundation for games played one-shot, like ours, and in experiments on other-regarding preferences (see Section 6.2 of Attanasi et al., 2016). We analyze the best-replies of A and B as a function, for the former, of her lexicographic altruism, and for the latter, of his lexicographic altruism and his guilt aversion.

2.3.2.1 Player A's Best-Reply Functions

As we do not use Bayesian equilibrium as a solution concept, and because we are mainly interested in B's behavior, here we only present a brief summary of A's best-reply analysis (the full analysis can be found in Appendix 2.B.1). The aim of this section is to show that given a (type, belief) pair of player A, her predicted behavior is game-dependent.

Table 2.3.1 summarizes, in each of the four Quasi-Trust mini-games, A's best-reply strategies as a function of: (i) her sensitivity to altruism toward B, ϕ_{AB}, in the Investment game, and toward C, ϕ_{AC}, in the other three games; (ii) her first-order belief that B

chooses *Right* after *In*, α_{AB}, for the four possible first-order beliefs about B choosing *Right* after *In* that A players can hold in our experiment, *i.e.*, $\alpha_{AB} \in \{0, 1/3, 2/3, 1\}$, as explained in Section 3.3.

Table 2.3.1: A's predicted behavior depending on her altruism sensitivity ϕ_{Ah} and first-order belief α_{AB}, with altruistic (resp., selfish) strategy in dark grey (resp., light grey) color.

Games	Investment				Rev. Investment				Donation				Exploitation			
α_{AB}	0	1/3	2/3	1	0	1/3	2/3	1	0	1/3	2/3	1	0	1/3	2/3	1
$\phi_{Ah} = 0$	*Out*	*Out*	*In*	*In*	*In*	*In*	*In*	*In*	*Out*	*Out*	*Out*	*Out*	*In*	*In*	*In*	*In*
$(0.00, 0.13)$	*Out*	*Out*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*Out*	*Out*	*Out*	*In*	*In*	*In*	*In*
$[0.13, 0.40)$	*Out*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*Out*	*Out*	*Out*	*In*	*In*	*In*	*In*
$[0.40, 0.50)$	*Out*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*Out*	*Out*	*Out*	*Out*	*In*	*In*	*In*
$[0.50, 0.68)$	*Out*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*Out*	*Out*	*In*	*Out*	*In*	*In*	*In*
$[0.68, 0.75)$	*Out*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*In*	*In*	*In*
$[0.75, 0.93)$	*In*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*In*	*In*	*In*
$[0.93, 1.07)$	*In*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*Out*	*In*	*In*
$[1.07, 1.47)$	*In*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*
$[1.47, 2.00)$	*In*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*In*	*In*	*In*	*Out*	*Out*	*Out*	*In*
$[2.00, 2.50)$	*In*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*	*Out*	*In*	*In*	*In*	*Out*	*Out*	*Out*	*Out*
$[2.50, +\infty)$	*In*	*In*	*In*	*In*	*Out*	*Out*	*In*	*In*	*In*	*In*	*In*	*In*	*Out*	*Out*	*Out*	*Out*

Strategy *In* is the altruistic one in all games but the Exploitation game, where *Out* is the altruistic strategy (see Figures 1-4). In the **Investment Game**, a selfish or lightly-altruistic A chooses *In* more, the higher is her first-order belief α_{AB} that B chooses *Right* after *In*; a highly-altruistic A chooses *In* regardless of α_{AB}. Thus, given α_{AB}, the higher ϕ_{AB}, the higher the likelihood that A chooses *In*. In the **Reversed-Investment Game**, a selfish A chooses *In* regardless of α_{AB}; any altruistic player A chooses *In* only for high enough α_{AB}. In the **Donation Game**, a selfish or lightly-altruistic A never chooses *In* regardless of α_{AB}; a highly-altruistic A chooses *In* only for high enough α_{AB}. In the **Exploitation Game**, a selfish or lightly-altruistic A always chooses *In* regardless of α_{AB}; a highly-altruistic A chooses *In* only for low enough α_{AB}.

Therefore, A's altruism leads to belief-dependent behavior. Furthermore, and more importantly, A's (belief-dependent) behavior is also game-dependent. In fact, the relation between altruism sensitivity ϕ_{Ah} and first-order belief α_{AB} which leads to *In* as a best-reply strategy differs across the four Quasi-Trust mini-games.

2.3.2.2 Player B's Best-Reply Functions

Recall that in each of the four Quasi-Trust mini-games if B chooses *Right* after *In*, he entirely transfers to another player the amount of money that A's *In* choice has entitled him to manage. If instead he chooses *Left* after *In*, he only transfers a small portion (20%) of that amount. Therefore, relying on Eqs. (2.3–2.5), for each treatment (A and C) we define B's *Willingness-to-Transfer function* (*WT*) as the difference between his utility from *Right* after *In* and his utility from *Left* after *In*. Both terms are expected utilities since B forms beliefs about the first-order beliefs α_{jB} of the co-player j toward whom he may feel guilty ($j = A$ in the treatment A and $j = C$ in the treatment C).[9] These are his conditional second-order beliefs $\beta_{Bj} = \mathbb{E}_B[\alpha_{jB}|In]$ for $j \in \{A, C\}$, *i.e.*, conditional on A choosing *In*.[10] The higher B's willingness to transfer the money that A's *In* choice has entitled him to manage, the more player B prefers to choose *Right* rather than *Left*:

$$\begin{aligned} WT(\phi_{BC}, \gamma_{Bjk}, \alpha_{jB}, z|In) &= \mathbb{E}_B[U_B(\phi_{BC}, \gamma_{Bjk}, \alpha_{jB}, R)] - \mathbb{E}_B[U_B(\phi_{BC}, \gamma_{Bjk}, \alpha_{jB}, L)] \\ &= \pi_B(R) - \pi_B(L) + \phi_{BC} \cdot [\pi_C(R) - \pi_C(L)] + \\ &\quad \gamma_{Bjk} \cdot \beta_{Bj} \cdot [\pi_k(R) - \pi_k(L)] \end{aligned} \tag{2.6}$$

[9] In each game, we assume that B best-responds *as if* he had truly observed A's move. This holds by standard expected-utility maximization, except for the cases where B is certain that A has chosen *Out*. Thus, we need the additional assumption that B has a belief conditional on *In*, even if he is certain of *Out*. Indeed, in our experiment (see Section 3.3) B's decision is made under the strategy method, *i.e.*, both when A has chosen *Out* and when she has chosen *In*.

[10] More precisely, we reason as if B has a point belief β_{Bj} about α_{jB} conditional on *In*.

More precisely, B chooses *Right* after *In* if $WT > 0$ in Eq. (2.6), and *Left* otherwise. Note that given their common structure, in each of the four games it is $\pi_B(R) - \pi_B(L) = -20$. With this, we can find player B's best-reply strategy as a function of his sensitivity to altruism toward player C, ϕ_{BC}, his second-order belief β_{Bj} that he will choose *Right* after *In*, and his sensitivity γ_{Bjk} to guilt toward the player j on whom B's second-order belief β_{Bj} relies.

In the **Investment Game**, B's strategy does not affect C's payoff thus B cannot be altruistic toward C (by construction, $F_{BC}(\phi_{BC}, L) = F_{BC}(\phi_{BC}, R)$ in Eq. (2.4)). Furthermore, B's strategy affects A's payoff, hence $k = A$ in Eq. (2.6). With this, WT in Eq. (2.6) reduces to:

$$\pi_B(R) - \pi_B(L) + \gamma_{BjA} \cdot \beta_{Bj} \cdot [\pi_A(R) - \pi_A(L)] \tag{2.7}$$

By substituting the game payoffs of Figure 2.2.1, Eq. (2.7) becomes $-20 + 40 \cdot \gamma_{BjA} \cdot \beta_{Bj}$, which is strictly positive for all type-belief pairs $(\gamma_{BjA}, \beta_{Bj})$ such that $\gamma_{BjA} \cdot \beta_{Bj} > 1/2$. Therefore, a guilt-averse B is more willing to choose *Right* after *In* for higher guilt sensitivity γ_{BjA} and higher second-order belief β_{Bj} of *Right* after *In*. This relationship holds both in treatment A, *i.e.*, for type-belief pairs $(\gamma_{BAA}, \beta_{BA})$, and in treatment C, *i.e.*, for $(\gamma_{BCA}, \beta_{BC})$.

In the **Reversed-Investment Game**, B's strategy affects C's payoff hence $k = C$ in Eq. (2.6), which becomes:

$$\pi_B(R) - \pi_B(L) + (\phi_{BC} + \gamma_{BjC} \cdot \beta_{Bj}) \cdot [\pi_C(R) - \pi_C(L)] \tag{2.8}$$

By substituting the game payoffs of Figure 2.2.2, Eq. (2.8) becomes $-20+40\cdot(\phi_{BC}+\gamma_{BjC}\cdot\beta_{Bj})$, which is strictly positive for all type-belief pairs $((\phi_{BC}, \gamma_{BjC}), \beta_{Bj})$ such that $\phi_{BC}+\gamma_{BjC}\cdot\beta_{Bj} > 1/2$. Therefore, a guilt-averse B is more willing to choose *Right* after *In* for higher guilt sensitivity γ_{BjC} and higher conditional second-order belief β_{Bj} of *Right* after *In*. This relationship holds both in treatment A, *i.e.*, for type-belief pairs $(\gamma_{BAC}, \beta_{BA})$, and in treatment C, *i.e.*, for $(\gamma_{BCC}, \beta_{BC})$. Furthermore, independently from the treatment, the higher ϕ_{BC}, B's sensitivity to altruism toward C, the lower both the guilt sensitivity γ_{BjC} and the second-order belief β_{Bj} required for B to choose *Right* after *In*. Finally, for high enough sensitivity to altruism (*i.e.*, $\phi_{BC} > 1/2$), player B chooses *Right* after *In* regardless of his second-order belief β_{Bj}.

In the **Donation Game**, B's strategy affects C's payoff, hence $k = C$ in Eq. (2.6). Therefore, WT in this game is the same as in Eq. (2.8). By substituting the game payoffs of Figure 2.2.3, given the similar structure between the Reversed-Investment and the Donation games ($\pi_B(R)-\pi_B(L) = -20$ and $\pi_C(R)-\pi_C(L) = 40$), we find the same subset of type-belief pairs $((\phi_{BC}, \gamma_{BjC}), \beta_{Bj})$ for which Eq. (2.8) is strictly positive, *i.e.*, $\phi_{BC}+\gamma_{BjC}\cdot\beta_{Bj} > 1/2$. Therefore, independently from treatment, the same considerations made for the Reversed-Investment game hold in the Donation game.

Finally, in the **Exploitation Game**, B's strategy does not C's payoff thus B cannot be altruistic toward C ($F_{BC} = 0$). Furthermore, B's strategy affects A's payoff , hence $k = A$ in Eq. (2.6). Therefore, WT in this game is the same as Eq. (2.7). By substituting the game payoffs of Figure 2.2.4, given the similar structure between the Investment and the Exploitation games ($\pi_B(R)-\pi_B(L) = -20$ and $\pi_A(R)-\pi_A(L) = 40$), we find the same subset of type-belief pairs $(\gamma_{BjA}, \beta_{Bj})$ for which Eq. (2.7) is strictly positive, *i.e.*, $\gamma_{BjA}\cdot\beta_{Bj} > 1/2$. Therefore, independently from treatment, the same considerations

made for the Investment game hold in the Exploitation game.

2.3.3 Hypotheses

2.3.3.1 Hypotheses on Player A's Behavior

Since we are mainly interested in B-subjects' behavior, we summarize briefly the hypotheses about A-subjects based on our model and we refer to Section 2.B.2 for details. Recall that in our experiment A-subjects are unaware of the treatment when they make their choices, hence A's behavior should be treatment-independent.

These hypotheses refer to two aspects of A's choices. First, H.A.1 and H.A.2 address A's lexicographic altruism in each game taken separately: as the theoretical predictions in Table 1 shows, a more trustful A-player is more willing to choose *In* regardless of the game, while the interplay between altruism sensitivity and willingness to choose *In* depends on the game. As for the latter, H.A.3 and H.A.4 specify A's motivation behind *In* choices across games. If these four hypotheses are supported, this would suggest that A's intention behind *In* is to increase C's payoff in the Reversed-Investment and Donation games, while she wishes to increase her own payoff in the Investment and Exploitation games.

H.A. 1. [Choice-belief correlation] *The frequency of In choices by A-subjects increases in their first-order belief about B-subjects choosing Right in each game.*

H.A. 2. [Choice-type correlation] *The frequency of In choices by A-subjects increases in their sensitivity to altruism in the Investment and the Donation games. It decreases*

in A-subjects' sensitivity to altruism in the Reversed-Investment and the Exploitation games.

H.A. 3. [Choice under beliefs of a distrustful A] *For A-subjects thinking that Left is the most likely action of B-subjects, the frequency of In choices in the Donation game is lower than: (i) in the Reversed-Investment game for selfish types; (ii) in the Exploitation game for selfish and lightly-altruistic types; (iii) in the Investment game for highly-altruistic types.*

H.A. 4. [Choice under beliefs of a trustful A] *For A-subjects thinking that Right is the most likely action of B-subjects, the frequency of In choices in the Investment game is: (i) the same as in the Reversed-Investment game, regardless of the altruistic type; (ii) higher than in the Donation game for selfish and lightly-altruistic types; (iii) higher than in the Exploitation game for highly-altruistic types.*

2.3.3.2 Hypotheses on Player B's Behavior

We have two families of hypotheses for B-subjects: a first one considering, in each game taken separately, correlations between B's choices and his second-order belief (H.B.1) or type (H.B.2); and a second one comparing B's decisions across games (H.B.3 to H.B.5).

Taken together, H.B.1 and H.B.2 postulate that guilt is activated in each of the eight game-treatment combinations. These hypotheses are at the core of our extension of Battigalli and Dufwenberg (2007). They contrast with predictions from Battigalli and Dufwenberg (2007) and follow-up studies that expect guilt to arise in only four treatment-game combinations where B's strategy and second-order beliefs are conditioned to the first-order beliefs of a player whose payoff depends on B's strategy (the Investment and Exploitation games in treatment A and the Reversed-Investment and Donation games in treatment C). Our contention in H.B.1 and H.B.2 is that B's guilt aversion is triggered

by his role of second mover in a two-stage game with perfect information: this role is game-independent, and so should be his sensitivity to guilt. [11]

H.B. 1. [Choice-belief correlation] *The frequency of Right choices by B-subjects increases in their second-order beliefs about Right in each of the four games.*

H.B. 2. [Choice-type correlation] *Given a positive second-order belief, the frequency of Right choices of B-subjects increases with: (i) their altruism sensitivity only in the Reversed-Investment and the Donation games; (ii) their guilt sensitivity in each of the four games.*

After assuming that B can feel guilty also when disappointing the beliefs of a player whose payoff is not affected by B's decision (H.B.1 and H.B.2), we now turn to the frequency of such guilt-averse behavior. Since our model is silent on this issue, we rely on Attanasi et al. (2019b) who tested this hypothesis in the Donation game and detected no significant difference in B's guilt between $j = k = C$ and $A = j \neq k = C$ in Eq. (2.3). Therefore, H.B.3 and H.B.4 posit the same fraction of guilt-averse B-players across the four games and across the two treatments.

H.B. 3. [Within-subject game-independent guilt] *Within a treatment, the fraction of guilt-averse B-subjects does not differ across the four games.*

H.B. 4. [Between-subject treatment-independent guilt] *Within a game, the fraction of guilt-averse B-subjects is not significantly different across treatments.*

[11] Note that, given A's first-order belief of *Right*, the same *In* choice in different games would signal an A's different sensitivity to altruism. Therefore, if B cares about the different intentions behind A's *In* choice, B's guilt sensitivity should also be game-dependent. However, our study relies on the opposite intuition that B's belief-dependent behavior is game-independent.

The joint test of H.B.3 and H.B.4 is the most original contribution of our study. Table 2.3.2 shows how this test helps us assessing whether (i) disappointing an ex-ante vulnerable player leads to higher guilt than disappointing a non-vulnerable one; (ii) disappointing an ex-post vulnerable player leads to higher guilt than disappointing a non-vulnerable one; (iii) disappointing an ex-post vulnerable player leads to higher or lower guilt than disappointing an ex-ante vulnerable one; (iv) disappointing an ex-ante and ex-post vulnerable player leads to higher guilt than disappointing a player vulnerable on just one of these two dimensions.

		Ex-post Vulnerability	
		No	Yes
Ex-Ante Vulnerability	No	A in Rev. Investment C in Investment	A in Exploitation C in Donation
	Yes	A in Donation C in Exploitation	A in Investment C in Rev. Investment

Table 2.3.2: Hypotheses on the proportion of guilt-averse B-players in each game-treatment combination

Notes: Game-treatment combinations in light grey (respectively, dark grey) represent situations where the partner is either ex-post or ex-ante vulnerable (respectively, both ex-post and ex-ante vulnerable).

Finally, relying on the assumption of B's lexicographic altruism, H.B.5 asserts that B's altruism is only activated in the Reversed-Investment and the Donation games. Since guilt aversion and altruism are the only other-regarding motivations of player B in Eq. (2.3), this would ultimately lead to a smaller fraction of selfish B-subjects (*i.e.*, always choosing *Left*) in these games.

H.B. 5. *[Game-dependent altruism] The fraction of B-subjects who behave selfishly is significantly higher in the Investment and the Exploitation games than in the Reversed-Investment and the Donation games. This holds independently of the treatment.*

2.4 Experimental Design and Procedures

In our experimental design, each subject went through the four Quasi-Trust games of Figures 2.2.1-2.2.4: Donation, Investment, Reversed-Investment and Exploitation. The games were renamed with neutral labels ("North", "South", "East", and "West"). In each game, subjects played in groups of three, with roles (A, B and C) assigned at the beginning of the session and maintained fixed across games. Groups were re-matched across games according to a perfect-stranger protocol. We randomized within-subjects the order of presentation of the four games across experimental sessions, and we varied between-subjects the treatments A and C.

2.4.1 Decisions and Elicitation of Beliefs

First-order belief elicitation We elicited, for each game, B-subjects' and C-subjects' first-order beliefs on the frequency of A-subjects choosing *In*. They had to report, for each game, their belief about the number of A-subjects, out of three A-subjects randomly selected in the session, who chose *In*, from 0 to 3 inclusive. We also elicited, for each game, A-subjects' and C-subjects' first-order beliefs on the frequency of B-subjects choosing *Right* after *In*. Similarly, they had to report, for each game, their belief about the number of B-subjects, out of three B-subjects randomly selected in the session, who chose *Right* conditional on A-subject choosing *In*, from 0 to 3 inclusive. For each role, one belief out of the four elicited in the four games was randomly selected at the end of the session and paid €1 if accurate.

A-subject's decision For each game, A-subjects chose between *In* or *Out*. At the end of the session, one of the four games was randomly selected to be payoff-relevant.

B-subject's decision B-subjects decided under the veil of ignorance, *i.e.*, assuming that their matched A-subject had chosen *In*. For each game, in treatment A (resp., treatment C) B-subjects made four decisions corresponding to their matched A-subject's (resp., C-subject's) four possible first-order beliefs on the frequency of *Right* choices conditional on *In*. In other words, in treatment A (resp., treatment C), B-subjects could condition their decision to the possible first-order beliefs of their matched A-subject (resp., C-subject). Given that the A-subject had chosen *In* in the game randomly selected to be payoff-relevant, the program implemented the B-subject's decision corresponding to the actual first-order belief of the A-subject (resp., C-subject) in treatment A (resp., treatment C). To facilitate decision making, the four possible first-order beliefs were presented in a fixed increasing order. This elicitation of decisions conditional on another player's first-order belief corresponds to the menu method of Khalmetski et al. (2015), which allows the experimenter to artificially induce second-order beliefs.[12]

Second-order belief elicitation We elicited, for each game, A-subjects' second-order beliefs on the frequency of A-subjects choosing *In* according to their matched B-subject's and C-subject's in the game. In other words, A had to guess B's and C's first-order beliefs on the frequency of A-subjects choosing *In*. We also elicited, for each game, B-subjects' second-order belief on the frequency of B-subjects choosing *Right* after *In* according to their matched A-subject and C-subject. Relying on previously elicited first-order beliefs, also second-order beliefs were elicited through asking subjects to report a number from

[12]The use of the menu method is now frequent in the experimental literature on guilt aversion (Khalmetski et al., 2015; Hauge, 2016; Balafoutas and Fornwagner, 2017; Bellemare et al., 2017; Dhami et al., 2017; Bellemare et al., 2018). Although one might argue that this method elicits "cold" responses, it offers several advantages. It allows to rule out potential false-consensus effects without raising the issue of strategic reporting and without using deception. The false-consensus effect could be avoided by communicating the A-subject's (C-subject's) true beliefs to B-subjects. However, it requires choosing between two evils: if A-subjects (C-subjects) know that their beliefs will be communicated, they are likely to distort them; and if they do not know that their beliefs will be communicated, the design is arguably deceptive. The menu method avoids these drawbacks. Moreover, it allows to study guilt aversion at the individual level and, hence, to unveil inter-individual differences that are hidden at the aggregate level (Khalmetski et al., 2015).

0 to 3 inclusive. For each role, one belief out of the four elicited (four games) was randomly selected at the end of the session and paid €1 if accurate.

2.4.2 Elicitation of Individual Preferences

In the second part of the experiment we elicited social preferences via the Social Value Orientation (SVO) test (Murphy et al., 2011). In the role of a decision maker, subjects made fifteen allocation choices between a decision maker and a passive player. They were paid for two randomly selected periods: one as a decision maker, one as a passive player.

Additionally, at the end of the session we collected non-incentivized measures of individual preferences, using the Guilt and Shame Proneness (GASP) questionnaire (Cohen et al., 2011). Moreover, subjects had to self-report their attitudes toward risk, patience and guilt proneness.[13] Finally, we collected socio-demographic characteristics, including gender, age, major and number of past participations in economic experiments.

2.4.3 Procedures

The experiment was conducted at GATE-Lab, Lyon, France. It was computerized using z-Tree (Fischbacher, 2007). Subjects were recruited mainly from the undergraduate

[13] Risk aversion and patience were measured by the following questions: “Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?" (Dohmen et al., 2011), and “Are you generally an impatient person, or someone who always shows great patience?" (Vischer et al., 2013). We adapted Moulton et al. (1966) to phrase the question on guilt proneness in a similar manner as for risk aversion and patience: “Are you generally a person who easily feel guilty or is it difficult to make you feel guilty?". Subjects rated how “easy" it is to make them feel guilty on a scale from 0 to 10, *i.e.*, with the same rating scale used to answer the two questions on how “willing to take risk" and how “patient" they are.

student population of local business, engineering and medical schools, using Hroot (Bock et al., 2014). 288 subjects participated in a total of 17 sessions. 57% were female and the average age was 22 years. Table 2.D.1 in Appendix 2.D shows that the mean individual characteristics are similar across treatments. Each session lasted about 75 minutes. Game payoffs were expressed in Experimental Currency Units (ECU) with 10 ECU = €1. Average earnings were €17 (S.D. = 5.91), including payment for accurate beliefs and a €5 show-up fee.

Upon arrival in the lab, subjects were randomly assigned to a cubicle after drawing a tag in an opaque bag. The session consisted of two parts. The instructions (Appendix 2.C) for the first part were distributed before each stage. The first stage described the four games. The experimenter made sure that all subjects had completed correctly a comprehension questionnaire before moving on to the second stage. At the beginning of the second stage, subjects were informed of their role. Then, we elicited the subjects' first-order beliefs and A-subjects made their decisions. In the third stage B-subjects made their decisions. Meanwhile, A- and C-subjects could solve sudoku-puzzles to avoid that their immediate neighbors in the lab could identify their role. In the fourth stage, we elicited the A- and B-subjects' second-order beliefs while C-subjects could solve sudoku puzzles. In the second part of the experiment, we implemented the SVO test. Then, subjects received feedback on their payoff and the decisions that were payoff-relevant, and they finally completed the socio-demographic questionnaire.

2.5 Results

Following the same logic as above, we first briefly summarize the main findings regarding A-subjects' behavior, with all details given in Appendix 2.D.2. Then, we focus on B-subjects.

2.5.1 A-Subjects' Behavior

As expected, the choice of *In* by the 96 A-subjects varies considerably across games. Pooling the two treatments, *In* is chosen by 48.87 % of the A-subjects in the Investment game, 70.83% in the Reversed-Investment game, 20.83% in the Donation game and 75% in the Exploitation game. We reject the null hypothesis that the proportion of A-subjects choosing *In* is the same across games (Cochran Q test; $p = 0.000$). Consistently, pairwise comparisons show that this proportion is significantly different across games (McNemar tests; highest $p = 0.001$ for Investment *vs.* Reversed-Investment), except when we compare the Reversed-Investment and the Exploitation games (70.83% *vs.* 75.00%; $p = 0.584$).[14]

We estimated separate Logit regressions for each game with the choice of *In* as the dependent variable (see also descriptive statistics in Table 2.D.2 in Appendix 2.D.2), and with their first-order beliefs and SVO angle as main independent variables (to test H.A.1 and H.A.2). These regressions pool the data from both treatments and control for the treatment, the order of the game and, according to the specification, for self-reported risk aversion and patience, and for socio-demographic variables (age, gender, number of previous experiments attended, and business major). They are reported in Table 2.D.3 in

[14]Except when specified otherwise, the non-parametric tests are two-sided and each decision is treated as one independent observation since only one decision per participant is payoff-relevant.

Appendix 2.D.2.[15]

Table 2.D.3 shows that the positive relationship between first-order belief and *In* choices is only supported in the Investment and the Donation games.

R.A. 1. *The frequency of* In *choices by A-subjects increases significantly in their first-order belief about B-subjects choosing* Right *in the Investment and the Donation games, but not in the Reversed-Investment and the Exploitation games.*

Table 2.D.3 supports the positive relationship between the SVO angle and *In* choices predicted by H.A.2 in the Investment and the Donation games, but it reports a non-significant relationship in the Reversed-Investment and the Exploitation games where it was predicted to be negative.

R.A. 2. *The frequency of* In *choices by A-subjects increases significantly in their altruism sensitivity in the Investment and the Donation games, but is not significantly influenced by their altruism sensitivity in the Reversed-Investment and the Exploitation games.*

To test H.A.3, we consider the choices of the A-subjects who believe that *Left* is the most likely action of B-subjects, *i.e.*, those with $\alpha_{AB} \leq 1/3$. We separate between selfish, lightly-altruistic and highly-altruistic A-subjects, as suggested by Table 2.3.1, by splitting them uniformly into these three categories according to their SVO angle. Precisely, we define as "selfish" the A-subjects with a SVO angle in the interval $(Min, Median - 15\%)$ of the empirical distribution, as "lightly-altruistic" those with a SVO angle in $(Median - 15\%, Median + 15\%)$, and as "highly-altruistic" those with

[15] Neither the treatment, nor the order of games have a significant effect on A-subjects' choices, except in the Investment game where treatment A significantly increases the frequency of *In* choices, at the 5% level. The results are robust to the inclusion of personality and socio-demographic controls.

a SVO angle in $(Median + 15\%, Max)$. The game differences predicted by H.A.3 are supported by this analysis, but note that most differences across games hold regardless of A-subjects' SVO angle.

R.A. 3. *For the A-subjects who believe that* Left *is the most likely action of B-subjects, the frequency of* In *choices in the Donation game is significantly lower than: (i) in the Reversed-Investment game for selfish types, (ii) in the Exploitation game for selfish and lightly-altruistic types, and (iii) in the Investment game for all altruistic types.*

Finally, to test H.A.4, we consider the choices of the A-subjects who believe that *Right* is the most likely action of B-subjects, *i.e.*, those with $\alpha_{AB} > 2/3$. We also separate among selfish, lightly-altruistic and highly-altruistic A-subjects. Most differences between the Investment game and the other games predicted by H.A.4 are supported by the analysis.

R.A. 4. *For the A-subjects who believe that* Right *is the most likely action of B-subjects, the frequency of* In *choices in the Investment game is: (i) not significantly different than in the Reversed-Investment game regardless of the type; (ii) higher, although not significantly so, than in the Donation game for selfish and lightly-altruistic types; (iii) significantly higher than in the Exploitation game for highly-altruistic types.*

2.5.2 B-Subjects' Behavior

We now explore B-subjects' behavior in depth. Before testing our hypotheses formally, we describe B-subjects' behavior through five patterns of choices for their four induced second-order beliefs, $\beta_{Bj} \in \{0, 1/3, 2/3, 1\}$, in each game:[16] (i) always choosing *Left*,

[16] By "induced second-order beliefs" we denote the four possible first-order beliefs of A-subjects (respectively, C-subjects) displayed on B-subjects' screens in treatment A (resp., treatment C).

regardless of the induced second-order beliefs, *i.e.*, choosing the payoff-maximizing (selfish) option (this represents on average 57% of the B-subjects);[17] (ii) always choosing *Right*, regardless of the induced second-order beliefs, *i.e.*, choosing the efficiency-maximizing option (5% of the B-subjects); (iii) switching from *Left* to *Right* as the induced second-order belief increases, *i.e.*, disclosing guilt aversion (26% of the B-subjects); (iv) switching from *Right* to *Left* as the induced second-order belief increases (6% of the B-subjects); and (v) any other pattern of choices (6% of the B-subjects). Figure 2.5.1 displays the distribution of B-subjects across these five patterns of choices in each game and for each treatment ($j = A$ and $j = C$) separately.[18]

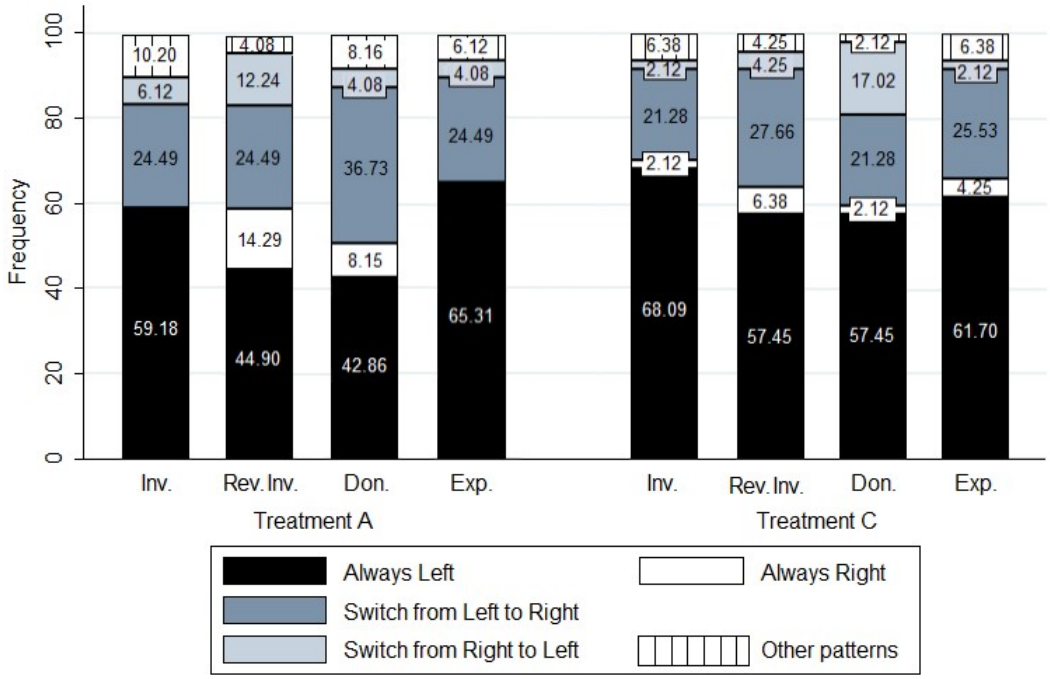


Figure 2.5.1: Distribution of B-Subjects' Pattern of Choices Across Games and Treatments

[17]The fact that the fraction of selfish B-subjects detected in our four games is on average higher than 50% is not surprising. Differently from the standard trust game, B's trustworthiness (*Right* if *In*) brings him no additional money with respect to his initial endowment, since $\pi_B(R) = \pi_B(O)$. Thus, in each game a B-subject choosing *Right* is purely driven by other-regarding preferences.

[18]In addition, Figure 2.D.1 in Appendix 2.D.3 analyzes the consistency of B-subjects' patterns of choices.

Note that among the five patterns of behavior identified below, our model is consistent with behaviors described in patterns (i) to (iii): they represent 87.54% of all B-subjects' behavior. More importantly, guilt-averse behavior (ii) represents 60% of all non-selfish behavior (patterns (ii) to (v)), thereby showing that guilt aversion is the prevailing social preference that is worth investigating in our games.

We now test our hypotheses and behavioral conjectures. Table 2.5.1 presents the marginal effects from panel Logit regressions on the probability to choose *Right*. For each game, we report two specifications. First, we regress B-subjects' choices on their induced and their stated second-order beliefs, β_{Bj}, their altruism sensitivity through their SVO angle, ϕ_{Bj}, and their self-reported guilt proneness, γ_{Bjk}. We control for the treatment and the order in which the game was played. The second specification adds personality (risk aversion and patience) and socio-demographic controls (age, gender, number of past participations in experiments, business major).

Table 2.5.1: Likelihood of B-Subjects Choosing *Right*, by Game

Game	Investment		Rev. Investment		Donation		Exploitation	
Induced SOB: β_{Bj}	0.186***	0.192***	0.187***	0.182***	0.197***	0.194***	0.208***	0.212***
	(0.045)	(0.045)	(0.048)	(0.047)	(0.050)	(0.048)	(0.045)	(0.044)
Stated SOB: β_{Bj}	0.270***	0.246***	0.434***	0.425***	0.411***	0.388***	0.245***	0.244***
	(0.075)	(0.072)	(0.087)	(0.081)	(0.096)	(0.084)	(0.064)	(0.059)
SVO Angle: ϕ_{Bj}	0.003*	0.002	0.010***	0.009***	0.004*	0.003	0.002	0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)
Reported Guilt: γ_{Bjk}	0.002	0.006	0.030***	0.033***	0.011	0.024**	0.005	0.013
	(0.008)	(0.008)	(0.011)	(0.011)	(0.011)	(0.010)	(0.008)	(0.009)
Treatment A	0.012	0.023	0.143**	0.120*	0.168***	0.181***	-0.032	-0.004
	(0.042)	(0.045)	(0.061)	(0.061)	(0.059)	(0.054)	(0.044)	(0.046)
Order	-0.013	-0.015	-0.010	-0.002	0.007	-0.014	-0.055**	-0.057***
	(0.023)	(0.025)	(0.025)	(0.024)	(0.022)	(0.022)	(0.023)	(0.022)
Personality	No	Yes	No	Yes	No	Yes	No	Yes
Socio-Demographics	No	Yes	No	Yes	No	Yes	No	Yes
N Observations	384	384	384	384	384	384	384	384
N subjects	96	96	96	96	96	96	96	96
Log-likelihood	-130.210	-126.656	-155.339	-152.365	-169.71	-159.82	-126.191	-120.278
Wald Chi2	28.82	31.69	36.24	38.12	32.30	40.39	31.78	34.36
Prob>chi2	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000

Notes: Table 2.5.1 reports the average marginal effects estimated by random-effects Logit models. Standard errors are in parentheses."Induced SOB β_{Bj}" corresponds to the four β_{Bj} presented to B-subjects when making their choice of *Right* or *Left*. "Stated SOB β_{Bj}" corresponds to the second-order beliefs reported by the B-subjects in the belief elicitation stage. "Reported Guilt" takes value between 0 and 10. "SVO angle" takes value between -7.8 and 38.9. "Order" is the rank order of the game, from 1 to 4. "Personality" controls correspond to the subjects' self-reported risk aversion and patience. "Socio-Demographics" controls include age, gender, number of previous experiments attended, business major. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

As predicted by H.B1, Table 2.5.1 shows that in all games, regardless of the specification, the higher is the supposed first-order belief of their matched A or C-subject (*i.e.*, B-subjects' induced second-order belief respectively in treatment A or C), the more likely B-subjects are to choose *Right*. The same holds for stated second-order beliefs: the likelihood of choosing *Right* significantly increases in B-subjects' stated second-order beliefs. Both are significant at the 1% level and regardless of the specification. Thus, H.B.1 is supported in each treatment and for any measure of second-order beliefs, as

stated in R.B.1.

R.B. 1. *The likelihood of B-subjects choosing* Right *significantly increases in their second-order beliefs about* Right *in each game, regardless of the treatment and for both induced and stated second-order beliefs.*

Regarding B-subjects' altruism sensitivity, Table 2.5.1 shows that, for a given second-order belief, a higher SVO angle increases significantly the likelihood of choosing *Right* in the Reversed-Investment game and not in the Investment or the Exploitation games, as predicted by H.B.2. In contrast, in the Donation game no effect reaches a standard significance level. It should be noted, however, that Spearman correlations between the SVO angle and the number of *Right* choices for the four induced second-order beliefs are significant and positive both in the Reversed-Investment and Donation games (respectively, $\rho_S = 0.42$, $p = 0.000$; $\rho_S = 0.30$, $p = 0.003$).

Regarding B-subjects' guilt sensitivity, Table 2.5.1 reveals a significant impact on the likelihood of choosing *Right* only in the Reversed-Investment and the Donation games, and in the latter only conditional on controlling for personality and socio-demographic characteristics.[19] With this, we conclude that H.B.2 is broadly supported for altruism sensitivity and only partially for guilt sensitivity, as stated in R.B.2.

R.B. 2. *The likelihood of B-subjects choosing* Right *increases in: (i) their altruism sensitivity in the Reversed-Investment game, and to some extent in the Donation game;*

[19]The absence of significance in the other two games may not be surprising given that, differently from the measure of altruism sensitivity, our measure of guilt sensitivity was not incentivized (see Bellemare et al., 2019, on the difficulty of finding empirical relationships between the concept of guilt aversion in economics and its characterization in psychological questionnaires).

(ii) their guilt sensitivity in the Reversed-Investment and the Donation games. This holds independently from the treatment.

Table 2.5.1 shows that treatment A has a positive effect on the likelihood of choosing *Right* in both the Reversed-Investment and the Donation games, the only ones where B's altruism is activated (see Eq. (5)). This can be due to an "example effect": in the games where A's *In* choices signal some other-regarding preferences (Donation and Reversed-Investment), B is more incline to be altruistic as well (more likely to choose Right), as if he was influenced by A's behavior. Also note that, for the Exploitation game, the later the game is presented to subjects, the less likely they are to choose *Right*, *i.e.*, disclosing other-regarding preferences. This game being the only one where A's *In* choice is uncontroversially selfish, the later it is presented, the more B-subjects have faced A's *In* choices that could be interpreted as less selfish, which may make B-subjects less likely to exhibit an other-regarding behavior in the Exploitation game.

H.B.3 states that guilt aversion is game-independent. To test this, we compare the proportion of B-subjects switching from *Left* to *Right*, *i.e.*, exhibiting guilt aversion, across games. Pooling the treatments, we cannot reject the null hypothesis that the proportion of guilt-averse B-subjects is the same across games (Cochran Q test; $p = 0.509$). Consistently, pairwise comparisons of games reveal no significant difference in the proportion of guilt-averse B-subjects (McNemar tests; lowest p-value, $p = 0.210$).[20] This holds for each treatment separately (lowest p-value within treatment A, $p = 0.109$; lowest p-value within treatment C, $p = 0.453$). H.B.3 is thus essentially supported, as

[20]Given our sample size, the odds ratio and a fixed error probability ($\alpha = 0.05$), we ran a post-hoc power analysis using G*Power (Faul et al., 2009). The highest power is achieved when comparing Investment *vs.* Donation: 21% ($\beta = 79\%$). By looking at achieved power as a function of sample size, we would need 563 B-subjects to obtain a power of 95%. The lowest power is achieved when comparing Reversed-Investment *vs.* Exploitation: 4% ($\beta = 96\%$). We would need 20 234 B-subjects to obtain a power of 95%.

stated in R.B.3.

R.B. 3. *The proportion of guilt-averse B-subjects is not significantly different across the four games, regardless of whether treatments are pooled together or not.*

H.B.4 states that the proportion of guilt-averse B-subjects in each game does not differ across treatments. To test this, we compare the proportion of B-subjects switching from *Left* to *Right* across treatments. Within a game, we find that the treatment has no significant impact on being guilt-averse in the Investment, the Reversed-Investment and the Exploitation games (Fisher exact tests; smallest $p = 0.810$ for the Investment game).[21] In the Donation game, the higher proportion of guilt-averse B-subjects in treatment A than in treatment C does not reach standard levels of significance (Fisher exact test; 36.73% *vs.* 21.28%; $p = 0.118$). This analysis confirms the importance of our extension of Battigalli and Dufwenberg (2007) to allow the emergence of guilt toward a player whose payoff does not depend on B's strategy. This finding is not in line with Bellemare et al. (2017) who detected more guilt among trustees than among dictators: in our three-player games, this should have translated into a higher guilt sensitivity of B-subjects toward a player signaling her intentions through her previous move (A, similar to a trustor in a trust game) than toward the passive C (similar to a recipient in a dictator game). Overall, this analysis supports H.B.4, as stated in R.B.4.

R.B. 4. *The proportion of guilt-averse B-subjects is not significantly different across treatments within each game.*

[21] Given our sample size, the odds ratio and a fixed error probability ($\alpha = 0.05$), a post-hoc power analysis shows that the highest power is achieved when comparing treatments in the Investment game, but only at the 5% level ($\beta = 95\%$). By looking at achieved power as a function of sample size, we would need 10198 B-subjects to obtain a power of 95%. The lowest power is achieved when comparing treatments in the Exploitation game: 4% ($\beta = 96\%$). We would need 96642 B-subjects to obtain a power of 95%.

We now turn to H.B.5, which predicts a higher proportion of selfish B-subjects, *i.e.*, those who always chose *Left*, in the Investment and the Exploitation games. Pooling the treatments, we indeed find that this proportion is higher in the Investment game than in the Reversed-Investment and Donation games (McNemar tests; $p = 0.012$ and $p = 0.007$, respectively). This proportion is also significantly higher in the Exploitation game than in the Reversed-Investment and Donation games ($p = 0.029$ and $p = 0.015$, respectively). These results hold if we consider treatment A separately (highest $p = 0.057$ for Investment *vs.* Donation game) but not in treatment C (lowest $p = 0.125$ for Investment *vs.* Donation game). We conclude that H.B.5 is mostly supported, as summarized in R.B.5.

R.B. 5. *B-subjects' probability of being selfish is significantly higher in the Investment and the Exploitation games than in the Reversed Investment and the Donation games both in treatment A and when treatments are pooled.*

Finally, following Bellemare et al. (2011) and Attanasi et al. (2019b), we define a structural econometric model to estimate B-subjects's average guilt sensitivity, γ_{Bjk}, toward player j's beliefs about player k's payoff, with $j,k \in \{A,C\}$ in each of the eight game-treatment combinations. Each B-subject chooses between *Right* and *Left* if *In* for each of the four possible first-order beliefs of j about *Right* if *In* (α_{jB}), in order to maximize his utility as defined by Eq. (2.9). In this Random Utility Model, λ is the noise parameter that we estimate and U_B essentially follows Eq. (2.3), for

$\alpha_{jB} \in \{0, 1/3, 2/3, 1\}$ and $j \in \{A, C\}$:[22]

$$V_B(\gamma_{Bjk}, \lambda, z|In) \quad = \quad U_B(\gamma_{Bjk}, z|In) + \lambda \cdot \varepsilon_B(z|In) \tag{2.9}$$

A conditional Logit model is used to estimate γ_{Bjk}, the sensitivity corresponding to B's guilt, $\max\{0, \mathbb{E}_j[\pi_k(z|In)] - \pi_k(z|In)\}$ in Eq. (2.5), while fixing to 1 the "sensitivity" corresponding to B's payoff $\pi_B(z|In)$. Table 2.5.2 reports the structural estimates of mean guilt sensitivity in each game-treatment combination, considering only B-subjects whose behavior is consistent with our model predictions (choosing always *Left* or always *Right* regardless of the four α_{jB}, or switching from *Left* to *Right* as α_{jB} increases – see Fig. 2.5.1). On average, they represent 87.54% of the B-subjects.

Table 2.5.2: Structural Estimates of Guilt sensitivity for B-Subjects Disclosing Behavior Consistent with the Model

	Treatment A				Treatment C				All
Game	Inv.	Rev-Inv.	Don.	Exp.	Inv.	Rev-Inv.	Don.	Exp.	
γ_{Bjk}	0.43*** (0.03)	0.45*** (0.06)	0.50*** (0.04)	0.39*** (0.04)	0.34*** (0.05)	0.39*** (0.04)	0.36*** (0.05)	0.36*** (0.05)	0.39*** (0.01)
N Obs.	328	328	344	352	304	344	344	344	2688

Pooling all games and treatments, we find that, on average, B-subjects are willing to pay 0.39 ECU to avoid disappointing their co-player's expectations by 1 ECU. Confidence intervals of the estimated γ_{Bjk} under the eight different specifications always overlap

[22] Recall that, differently from Bellemare et al. (2011), in our model B can also be altruistic toward C, with altruism sensitivity measured through the parameter ϕ_{BC} (see Eq. (2.4)). However, the second component of B's feeling of altruism F_{BC}, *i.e.*, C's material payoff π_C, is colinear with B's material payoff π_B by design. Therefore, we cannot estimate the three coefficients (ϕ_{BC}, γ_{Bjk}, and the coefficient corresponding to π_B) of our utility function while estimating the noise parameter of our random utility model in Eq. (2.9). Thus, we renounce to estimate ϕ_{BC} in the two games where it is assumed to be non-null, *i.e.*, the Reversed-Investment and the Donation games. In the remaining games, this is not an issue since by design $\phi_{BC} = 0$ (lexicographic altruism).

with the confidence interval of our benchmark (All). There is no combination of game and treatment where the desire to avoid disappointing a co-player is higher or lower than the average. More generally, when comparing estimated γ_{Bjk} by game and treatment, confidence intervals always overlap. Therefore, these structural estimates essentially confirm R.B.3 and R.B.4.

2.6 Conclusion

Using four three-player Quasi-Trust mini-games, the current study identified different types of players' vulnerability as potential factors influencing a second-mover's guilt toward the other two players. We found that neither the proportion of guilt-averse second movers nor the intensity of their guilt aversion differed significantly across the four games (*i.e.* the four combinations of vulnerability), and across the two treatments (*i.e.* guilt elicited toward an active *vs.* a passive player). In particular, second movers exhibited a guilt-averse behavior even toward the beliefs of players who were not vulnerable at all. In doing so, we revealed the relevance of guilt aversion in games where it had never been tested before (for an exception, see Attanasi et al., 2019b, Chapter 1).

The main contribution of the present study is to evidence, both empirically and theoretically, the independence of guilt aversion from the vulnerability of the decision-maker's co-players. Theoretically, the current study develops a model where guilt aversion depends neither on the game, nor on the treatment, but rather depends on the role played by the decision maker. We contend that guilt is activated even when the beliefs of the disappointed player do not concern her material payoff but the payoff of a third player; this is a crucial assumption of our model, as in Attanasi et al. (2019b) (Chapter 1).[23]. This

[23] Attanasi et al. (2019b) were the first to show that "guilt towards another player can be triggered even when decisions have no direct consequences for that player" (Dufwenberg and Patel (2019, p. 3))

can be interpreted as further support to guilt mainly being role-dependent in two-stage games with asymmetric roles.

Our secondary objective was to assess the impact of the co-players' status (active *vs.* passive) on the decision-maker's guilt aversion. We showed that the second mover's guilt aversion was triggered even though the first mover's intention was mediated by a passive player. This result suggests that observing the intentions of co-players is not a necessary condition to trigger guilt.

The insensitivity of the second-mover's guilt aversion to manipulations of the co-player's vulnerability and intentions could, however, be interpreted as a sign of confusion in our subjects. Yet, the first movers' behavior pleads against this interpretation. We found their behavior to be game-dependent, in line with our model of lexicographic-altruism where we assume that the first-mover can feel altruism toward the second mover in the Investment game (where the passive player is a simple observer) and toward the passive player in the other three games. Alternatively, the way we elicit B-subjects' choices may have reduced the potential impact of the co-players' vulnerability. Indeed, participants were asked to make their choice conditional on four levels of expectations of the other player. This contextualization of choices, traditional when testing belief-based preferences (see, *e.g.*, Khalmetski et al., 2015), has potentially overcome the information provided when introducing the game that was supposed to trigger a reaction based on the other player's vulnerability. This alternative account of our results could be tested by first informing B-subjects of the co-player's expectations and then asking them to condition their choices on the different manipulations of their co-players' vulnerability.

Bibliography

Andrighetto, G., Grieco, D., and Tummolini, L. (2015). Perceived legitimacy of normative expectations motivates compliance with social norms when nobody is watching. *Frontiers in Psychology*, 6:1413.

Attanasi, G., Battigalli, P., and Manzoni, E. (2016). Incomplete-information models of guilt aversion in the trust game. *Management Science*, 62(3):648–667.

Attanasi, G., Battigalli, P., Manzoni, E., and Nagel, R. (2019a). Belief-dependent preferences and reputation: Experimental analysis of a repeated trust game. *Journal of Economic Behavior & Organization*, 167:341–360.

Attanasi, G., Battigalli, P., and Nagel, R. (2013). Disclosure of belief-dependent preferences in a trust game. Technical report, No. 506, IGIER (Innocenzo Gasparini Institute for Economic Research), Bocconi University.

Attanasi, G., Rimbaud, C., and Villeval, M. C. (2019b). Embezzlement and guilt aversion. *Journal of Economic Behavior & Organization*, 167:409–429.

Balafoutas, L. and Fornwagner, H. (2017). The limits of guilt. *Journal of the Economic Science Association*, 3(2):137–148.

Balafoutas, L. and Sutter, M. (2017). On the nature of guilt aversion: Insights from a new methodology in the dictator game. *Journal of Behavioral and Experimental Finance*, 13:9–15.

Battigalli, P. and Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, 97(2):170–176.

Battigalli, P. and Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory*, 144(1):1–35.

Battigalli, P. and Dufwenberg, M. (2020). Belief-dependent motivations and psychological game theory.

Baumeister, R. F., Stillwell, A. M., and Heatherton, T. F. (1994). Guilt: an interpersonal approach. *Psychological bulletin*, 115(2):243.

Bellemare, C., Sebald, A., and Strobel, M. (2011). Measuring the willingness to pay to avoid guilt: estimation using equilibrium and stated belief models. *Journal of Applied Econometrics*, 26(3):437–453.

Bellemare, C., Sebald, A., and Suetens, S. (2017). A note on testing guilt aversion. *Games and Economic Behavior*, 102:233–239.

Bellemare, C., Sebald, A., and Suetens, S. (2018). Heterogeneous guilt sensitivities and incentive effects. *Experimental Economics*, 21(2):316–336.

Bellemare, C., Sebald, A., and Suetens, S. (2019). Guilt aversion in economics and psychology. *Journal of Economic Psychology*, 73:52–59.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122–142.

Bock, O., Baetge, I., and Nicklisch, A. (2014). hroot: Hamburg registration and organization online tool. *European Economic Review*, 71:117–120.

Bolton, G. E. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American economic review*, 90(1):166–193.

Bracht, J. and Regner, T. (2013). Moral emotions and partnership. *Journal of Economic Psychology*, 39:313–326.

Buskens, V. and Raub, W. (2013). *Rational choice research on social dilemmas: embeddedness effects on trust*. Russell Sage: New York, NY, USA.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6):1579–1601.

Cohen, T. R., Wolf, S. T., Panter, A. T., and Insko, C. A. (2011). Introducing the gasp scale: a new measure of guilt and shame proneness. *Journal of personality and social psychology*, 100(5):947.

Cox, J. C., Kerschbamer, R., and Neururer, D. (2016). What is trustworthiness and what drives it? *Games and Economic Behavior*, 98:197–218.

Danilov, A., Khalmetski, K., and Sliwka, D. (2019). Descriptive norms and guilt aversion. *Mimeo*.

Dhami, S., Wei, M., and al Nowaihi, A. (2017). Public goods games and psychological utility: Theory and evidence. *Journal of Economic Behavior & Organization*.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3):522–550.

Dufwenberg, M. and Patel, A. (2019). Introduction to special issue on psychological game theory. *Journal of Economic Behavior & Organization*, 167(C)(3):181–184.

Engler, Y., Kerschbamer, R., and Page, L. (2018). Why did he do that? using counterfactuals to study the effect of intentions in extensive form games. *Experimental Economics*, 21(1):1–26.

Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2009). Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3):817–868.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, 10(2):171–178.

Geanakoplos, J., Pearce, D., and Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and economic Behavior*, 1(1):60–79.

Hauge, K. E. (2016). Generosity and guilt: The role of beliefs and moral standards of others. *Journal of Economic Psychology*, 54:35–43.

Kawagoe, T. and Narita, Y. (2014). Guilt aversion revisited: An experimental test of a new model. *Journal of Economic Behavior & Organization*, 102:1–9.

Khalmetski, K. (2016). Testing guilt aversion with an exogenous shift in beliefs. *Games and Economic Behavior*, 97:110–119.

Khalmetski, K., Ockenfels, A., and Werner, P. (2015). Surprising gifts: Theory and laboratory evidence. *Journal of Economic Theory*, 159:163–208.

Moulton, R. W., Burnstein, E., Liberty Jr, P. G., and Altucher, N. (1966). Patterning of parental affection and disciplinary dominance as a determinant of guilt and sex typing. *Journal of Personality and Social Psychology*, 4(4):356.

Murphy, R. O., Ackermann, K. A., and Handgraaf, M. (2011). Measuring social value orientation. *Judgment and Decision making*, 6(8):771–781.

Pelligra, V., Reggiani, T., and Zizzo, D. J. (2020). Responding to (un) reasonable requests by an authority. *Theory and Decision*, pages 1–25.

Vischer, T., Dohmen, T., Falk, A., Huffman, D., Schupp, J., Sunde, U., and Wagner, G. G. (2013). Validating an ultra-short survey measure of patience. *Economics Letters*, 120(2):142–145.

Appendix

2.A Literature

Table 2.A.1 presents a lists of published papers, citing Battigalli and Dufwenberg (2007) with an explicit reference to guilt aversion as a motivation of behavior and including an experiment.[24] It shows that 53.84% of the papers corresponds to trust games while 30.76% corresponds to dictator games. Hence, it also means that only 15.38% of the literature on guilt aversion has investigated other games.

[24]This list was compiled based on the authors knowledge of the literature.

Article	Game
Vanberg (2008)	Dictator
Reuben et al. (2009)	Trust
Ellingsen et al. (2010)	Dictator
Bellemare et al. (2011)	Trust
Chang et al. (2011)	Trust
Charness and Dufwenberg (2011)	Participation
Dufwenberg et al. (2011)	Coordination
Pelligra (2011)	Trust
Attanasi et al. (2013)	Trust
Amdur and Schmick (2013)	Trust
Battigalli et al. (2013)	Sender-Receiver
Beck et al. (2013)	Credence Good
Bracht and Regner (2013)	Trust
Kawagoe and Narita (2014)	Trust
Ockenfels and Werner (2014)	Dictator
Regner and Harth (2014)	Trust
Andrighetto et al. (2015)	Trust
Khalmetski et al. (2015)	Dictator
Yu et al. (2015)	Trust
Attanasi et al. (2016)	Trust
Hauge (2016)	Dictator
Ismayilov and Potters (2016)	Trust
Khalmetski (2016)	Sender-Receiver
Balafoutas and Sutter (2017)	Dictator
Balafoutas and Fornwagner (2017)	Dictator
Bellemare et al. (2017)	Trust & Dictator
Dhami et al. (2017)	Public Good
Ederer and Stremitzer (2017)	Dictator
Bellemare et al. (2018)	Dictator
Engler et al. (2018)	Trust
Attanasi et al. (2019a)	Trust
Attanasi et al. (2019b)	Embezzlement
Bellemare et al. (2019)	Trust & Dictator
Di Bartolomeo et al. (2019)	Trust
Inderst et al. (2019)	Trust
Morell (2019)	Dictator
Ciccarone et al. (2020)	Dictator
Ghidoni and Ploner (2020)	Lost Wallet
Peeters and Vorsatz (2021)	Prisoner Dilemma

Table 2.A.1: List of published experiments on guilt aversion

2.B Player A's Best-Reply Functions and Hypotheses

2.B.1 Player A's Best-Reply Functions

In each game, A's best-reply strategy is defined as a function of her first-order belief α_{AB} that B chooses *Right* after *In* and of her sensitivity to altruism toward B, ϕ_{AB}, in the Investment game, and toward C, ϕ_{AC}, in the other three games.

By construction, in each game A's altruism toward player h depends on her belief about B's action after *In*. Let us define the net expected altruism of a player A with altruistic sensitivity ϕ_{Ah} toward player h. It is the difference between her expected altruism when she chooses *In* and her altruism when she chooses *Out*, where α_{AB} is A's first-order belief that B chooses *Right* after *In*:

$$\mathbb{E}_A[F_{Ah}(\phi_{Ah}, z|In)] - F_{Ah}(\phi_{Ah}, Out) = \phi_{Ah} \cdot [\alpha_{AB} \cdot \pi_h(R) + (1-\alpha_{AB}) \cdot \pi_h(L) - \pi_h(O)] \quad (2.10)$$

Investment Game. In this game, choosing *In* does not affect C's material payoff while it affects B's expected material payoff. Therefore, by lexicographic altruism, A is altruistic toward B. Relying on A's expected altruism toward B (Eq. (2.10) with $h = B$), we conclude that A chooses *In* if $20 \cdot \phi_{AB} \cdot (1-\alpha_{AB}) \geq 0$. Relying on A's expected material payoff, we conclude that A chooses *In* if $\alpha_{AB} \cdot 195 + (1-\alpha_{AB}) \cdot 155 - 170 \geq 0$, *i.e.*, if $\alpha_{AB} \geq 3/8$. Putting together A's material and altruistic interest, we find that A chooses *In* if:

$$\phi_{AB} \geq \frac{3 - 8 \cdot \alpha_{AB}}{4(1-\alpha_{AB})} \quad (2.11)$$

From Eq. (2.11) follows that if $\alpha_{AB} \geq 3/8$, then A chooses *In* whatever her altruism sensitivity ϕ_{AB}. For lower first-order beliefs α_{AB}, A chooses *In* only if she is altruistic ($\phi_{AB} > 0$), where the lower the α_{AB}, the higher the altruism sensitivity needed to choose *In*. In particular, a highly-altruistic A ($\phi_{AB} \geq 3/4$) chooses *In* regardless of her first-order belief α_{AB}.

Reversed-Investment Game. In this game, choosing *In* affects C's material payoff. Therefore, by lexicographic altruism, A is altruistic toward C. Relying on A's expected altruism toward C (Eq. (2.10) with $h = C$), we conclude that A chooses *In* if:

$$\phi_{AC} \cdot (40 \cdot \alpha_{AB} - 15) \geq 0 \tag{2.12}$$

i.e., if $\alpha_{AB} \geq 3/8$. In this game, choosing *In* does not affect A's material payoff. Therefore, A's best reply function relies only on Eq. (2.12). With this, if $\alpha_{AB} \geq 3/8$, any altruistic A (*i.e.*, with $\phi_{AC} > 0$) chooses *In* regardless of her sensitivity to altruism, as in the Investment Game. However, differently from the Investment Game, any altruistic A chooses *Out* for $\alpha_{AB} < 3/8$. Eq. (2.12) also shows that a selfish A ($\phi_{AC} = 0$) is indifferent between *In* and *Out* for each first-order belief α_{AB}: we assume that she chooses *In*. The way we break the tied strategies is motivated by experimental demand effects due to both welfare maximization (see, e.g., Charness and Rabin, 2002) and to the fact that choosing *In* let the game unfold with B's strategy being payoff-relevant for himself and player C. Note that this assumption applies as tie breaking rule in all indifferences in Eqs. (2.11-2.14).

Donation Game. In this game choosing *In* affects C's material payoff. Therefore, by lexicographic altruism, A is altruistic toward C. Relying on A's expected altruism toward C (Eq. (2.10) with $h = C$), we conclude that A chooses *In* if $\phi_{AC} \cdot (40 \cdot \alpha_{AB} + 10) \geq 0$. Relying only on A's expected material payoff, we conclude that A never chooses *In* since $-25 < 0$. Putting together A's material and altruistic interest, we find that A chooses *In* if:

$$\phi_{AC} \geq \frac{5}{2(1 + 4 \cdot \alpha_{AB})} \tag{2.13}$$

Thus, a necessary condition for choosing *In* is altruistic enough toward player C, *i.e.*, $\phi_{AC} \geq 1/2$. But this is not sufficient: A's first-order belief of *Right* after *In* must be high enough, with higher ϕ_{AC} compensating for lower α_{AB}. At the limit, for $\alpha_{AB} = 0$, only A's types with $\phi_{AC} \geq 5/2$ choose *In*. Thus, the best-reply behavior of A's (type, belief) pairs in this game is qualitatively similar

to the one in the Investment Game, featuring a positive type-belief interaction. However, given A's altruistic type (resp., belief of *Right* after *In*) in both games, a higher belief of *Right* after *In* (resp., altruistic type) is needed to choose *In* in the Donation Game.

Exploitation Game. In this game choosing *In* affects C's material payoff. Therefore, by lexicographic altruism, A is altruistic toward C. Relying on A's expected altruism toward C (Eq. (2.10) with $h = C$), we conclude that A chooses *In* if $-25 \cdot \phi_{AC} \geq 0$. Hence, an altruistic A ($\phi_{AC} > 0$) never chooses *In*: differently from the other three games, the altruistic action is *Out*. Relying on A's expected material payoff, we conclude that A chooses *In* if $40 \cdot \alpha_{AB} + 10 > 0$. Putting together A's material and altruistic interest, we find that A chooses *In* if:

$$\phi_{AC} \leq \frac{2(1+4 \cdot \alpha_{AB})}{5} \tag{2.14}$$

Thus, a necessary condition for choosing *In* is that A is not too altruistic toward player C, *i.e.*, $\phi_{AC} < 2$. But this is not sufficient: A's first-order belief of *Right* after *In* must be high enough, with lower ϕ_{AC} compensating for lower α_{AB}. At the limit, for $\alpha_{AB} = 0$, only A's types with $\phi_{AC} \leq 2/5$ choose *In*. Thus, the best-reply relation between A's type and A's belief in this game is of opposite sign of the one in the Donation Game: given A's belief of *Right* after *In*, a lower altruistic type is needed to choose *In*.

All of the above is summarized in Table 2.3.1 of Section 3.2.1.

2.B.2 Player A's Behavioral Hypotheses

Given A's sensitivity to altruism, the theoretical predictions in Table 2.3.1 show a positive relationship between the likelihood of the *In* choice and A's first-order belief of *Right* after *In*. This holds regardless of the game. In particular, in the Investment game it holds regardless of A's sensitivity to altruism, in the Reversed-Investment game for all altruistic subjects, in the Donation game only for highly-altruistic subjects, and in the Exploitation game for all subjects but highly-altruistic

ones. Considering heterogeneity in A's sensitivity to altruism, we elaborate an hypothesis about A's belief-dependent behavior in each game.

H.A.1. [Choice-belief correlation] The frequency of *In* choices by A-subjects increases in their first-order belief about B-subjects choosing *Right* in each game.

Given A's first-order belief of *Right* after *In*, the theoretical predictions in Table 2.3.1 for the Investment and the Donation games show a positive relationship between the likelihood of the *In* choice and A's sensitivity to altruism. For the Investment game, this only holds under first-order beliefs of a distrustful B ($\alpha \leq 1/3$). For the Donation game, the positive relationship holds for every positive first-order belief. Conversely, for both the Reversed-Investment and the Exploitation games they show a negative relationship between the likelihood of the *In* choice and A's sensitivity to altruism. For the Reversed-Investment game, this only holds under first-order beliefs of a distrustful B ($\alpha \leq 1/3$). For the Exploitation game, the negative relationship holds for every first-order belief. Considering heterogeneity in both A's sensitivity to altruism and A's first-order belief of *Right* after *In*, the second hypothesis about A's altruistic type-dependent behavior is as follows:

H.A.2.[Choice-type correlation] The frequency of *In* choices by A-subjects increases in their sensitivity to altruism in both the Investment and the Donation games. It decreases in A-subjects' sensitivity to altruism in both the Reversed-Investment and the Exploitation games.

In the next two hypotheses, besides heterogeneity in A's sensitivity to altruism, we also make the operational assumption that this sensitivity does not vary too much across games within-subject (*i.e.*, each A-subject has a ϕ_{Ah} in the same interval of Table 2.3.1 for each of the four games). This is required in order to elaborate between-game comparisons in terms of type-dependent behavior, given a low ($\alpha_{AB} \leq 1/3$) or a high ($\alpha_{AB} \geq 2/3$) first-order belief about B choosing *Right*, *i.e.*,

given beliefs of a distrustful B or a trustful B, respectively.

For A's choice under beliefs of a distrustful B, we elaborate the next hypothesis by looking at the first two columns ($\alpha_{AB} \in \{0, 1/3\}$) of each of the four game panels of Table 2.3.1. We take the Donation game as the reference (control) since it is the only one where it is possible to identify a subset of altruistic A-types predicted to choose *Out* in that game and *In* in at least another game. In fact, for $\alpha_{AB} = 0$, A chooses *Out* in the Donation game regardless of $\phi_{Ah} < 2.5$, whereas in the other three games there exists a subset of A-types with $\phi_{Ah} < 2.5$, predicted to choose *In*.[25] These types are A-subjects with $\phi_{Ah} \geq 0.75$ in the Investment Game, with $\phi_{Ah} = 0$ in the Reversed-Investment Game, and with $\phi_{Ah} < 0.40$ in the Exploitation Game. Table 2.3.1 shows a similar pattern for $\alpha_{AB} = 1/3$: A-types with $\phi_{Ah} \geq 0.13$, with $\phi_{Ah} = 0$ and with $\phi_{Ah} < 0.93$ are predicted to choose *In* respectively in the Investment, Reversed-Investment and Exploitation Game, but not in the Donation Game. We combine the two sets of predictions for $\alpha_{AB} = 0$ and $\alpha_{AB} = 1/3$ in a unique hypothesis about game-dependent behavior of A-subjects who believe that B would be distrustful after *In*, *i.e.*, that he would more likely choose *Left* ($\alpha_{AB} \leq 1/3$)).

H.A.3. [Choice under beliefs of a distrustful A] For A-subjects thinking that *Left* is the most likely action of B-subjects, the frequency of *In* choices in the Donation game is lower than: (i) in the Reversed-Investment game for selfish types; (ii) in the Exploitation game for selfish and slightly-altruistic types; (iii) in the Investment game for slightly-altruistic and highly-altruistic types.

For A's choice under beliefs of a trustful B ($\alpha_{AB} \geq 2/3$), we elaborate the next hypothesis by looking at the last two columns ($\alpha_{AB} \in \{2/3, 1\}$) of each of the four game panels of Table 2.3.1. We take the Investment game as the reference (control) since it is the only one where it is possible

[25] We are aware from extensive experimental literature eliciting sensitivity to altruism that subjects with $\phi_{Ah} \geq 2.5$ are quite rare in the population. They would be indifferent between keeping 2.5 euros to themselves and giving 1 euro to another player (see, e.g., Andreoni et al., 2010; Bellemare et al., 2008). That is why elaborating H.A.3 for $\phi_{Ah} < 2.5$ is without loss of generality.

to identify a subset of A-types predicted to choose *In* in that game and *Out* in at least another game. In fact, A chooses *In* in the Investment Game regardless of ϕ_{Ah}, whereas in the two of the other games there exists a subset of A-types predicted to choose *Out*. For $\alpha_{AB} = 2/3$, these types are A-subjects with $\phi_{Ah} \geq 1.47$ in the Exploitation Game and with $\phi_{Ah} < 0.68$ in the Donation Game. For $\alpha_{AB} = 1$, these types are A-subjects with $\phi_{Ah} \geq 2.00$ in the Exploitation Game and with $\phi_{Ah} < 0.5$ in the Donation Game. We combine the two sets of predictions for $\alpha_{AB} = 2/3$ and $\alpha_{AB} = 1$ in a unique hypothesis about game-dependent behavior of A-subjects who believe that B would be trustful after *In*, *i.e.*, that he would more likely choose *Right* ($\alpha_{AB} > 1/3$)).

H.A.4. [Choice under beliefs of a trustful A] For A-subjects thinking that *Right* is the most likely action of B-subjects, the frequency of *In* choices in the Investment game is: (i) the same as in the Reversed-Investment game, regardless of the altruistic type; (ii) higher than in the Donation game for selfish and slightly-altruistic types; (iii) higher than in the Exploitation game for highly-altruistic types.

Note that we derived 1 to 4 without specifying the treatment (A or C) since, in our experiment, players A are unaware, when they make their choices, of the treatment. Therefore, A's behavior should be treatment-independent.

2.C Instructions (Translated from French)

We thank you for participating in this experimental session on decision-making. During this session, you can earn money. The amount of your earnings depends both on your decisions and on the decisions of other participants. At the end of the session, you will receive your earnings in cash in a separate room to preserve the confidentiality of your earnings. The earnings you will receive will include:

- your earnings from today's session
- a €5 fee for showing-up on time to the session.

During the session, some of the transactions are conducted in ECU (Experimental Currency Units).

Please turn off your phone. Communication with the other participants is prohibited during the entire duration of the session. If you have questions during the session, raise your hand or press the red button on the side of your desk and we will come to answer in private.

OVERVIEW OF THE SESSION

In this session, there are two parts. The two parts are completely independent. In each part, one or more of your decisions will be randomly selected by the computer. At the end of the session, you will be informed of your decisions, the decisions of other participants (if they affect your earnings) and their impact on your earnings.

At the end of the session you will be asked to answer a final questionnaire.

FIRST PART: OVERVIEW

In this part, the conversion rate is as follows: 10 ECU = €1.

Roles: At the beginning of the first part, the computer program randomly assigns a role to each participant. You can be either a participant A, a participant B or a participant C. Your role is indicated on your computer screen at the beginning of the first part and you keep the same role throughout this part.

Then, the computer program randomly forms groups of three participants, with one participant of each role in each group. The computer program forms a new group for each situation (which we will describe below), so your group composition changes during the first part. You will never know the

identity of the other members of your group and they will never be informed on your identity.

Decisions: Each participant receives an initial endowment. First, Participant A has to make a decision. He can send 25 ECU to Participant B or not. The 25 ECU sent to Participant B come from the endowment of either Participant A or Participant C, depending on the situation.

Then, if Participant B has received 25 ECU, he has to make a decision. He decides how to distribute these 25 ECU between another participant (A or C, depending on the situation) and himself. The ECU that Participant B transfers to another participant (A or C, depending on the situation) are multiplied by two, whereas the ECU that Participant B keeps for himself are not multiplied by two.

Situations: There are four different situations: "North", "West", "East" and "South" (the name of each situation has been given arbitrarily). Decisions are made in each of these four situations.

- In the North situation, Participant A decides whether or not to send 25 ECU from his initial endowment to Participant B. If Participant B receives these 25 ECU, he decides how to distribute these 25 ECU between Participant C and himself.
- In the West situation, Participant A decides whether or not to send 25 ECU of his initial endowment to Participant B. If Participant B receives these 25 ECU, he decides how to distribute these 25 ECU between Participant A and himself.
- In the East situation, Participant A decides whether or not to send 25 ECU from the initial endowment of Participant C to Participant B. If Participant B receives these 25 ECU, he decides how to distribute these 25 ECU between Participant C and himself.
- In the South situation, Participant A decides whether or not to send 25 ECU from the initial endowment of Participant C to Participant B. If Participant B receives these 25 ECU, he decides how to distribute these 25 ECU between Participant A and himself.

We will now describe in details the roles, decisions and situations in the first part.

FIRST PART: ROLES, DECISIONS, SITUATIONS

Participant A receives an initial endowment of 170 ECU.

He decides whether or not to send 25 ECU from either his endowment or Participant C's endowment to Participant B.

In the North situation, Participant A has the choice between:

- sending 25 ECU from his initial endowment to Participant B
- sending 0 ECU from his initial endowment to Participant B

In the West situation, Participant A has the choice between:

- sending 25 ECU from his initial endowment to Participant B
- sending 0 ECU from his initial endowment to Participant B

In the East situation, Participant A has the choice between:

- sending 25 ECU from Participant C's initial endowment to Participant B
- sending 0 ECU from Participant C's initial endowment to Participant B

In the South situation, Participant A has the choice between:

- sending 25 ECU from Participant C's initial endowment to Participant B
- sending 0 ECU from Participant C's initial endowment to Participant B

Participant B receives an initial endowment of 100 ECU.

If Participant A has sent 25 ECU to Participant B, Participant B has to make a decision. Then, participant B decides how to distribute these 25 ECU between another participant (A or C, depending on the situation) and himself. The ECU that Participant B transfers to another participant (A or C, depending on the situation) are doubled, whereas the ECU that Participant B keeps for himself are not doubled.

In the North situation, Participant B has the choice between:

- transferring the 25 ECU to the participant C - the participant C receives 50 ECU

- transferring 5 ECU to the participant C - the participant C receives 10 ECU - and keeping 20 ECU for himself - the participant B keeps 20 ECU.

In the West situation, Participant B has the choice between:

- transferring the 25 ECU to Participant A - Participant A receives 50 ECU
- transferring 5 ECU to Participant A - Participant A receives 10 ECU - and keeping 20 ECU for himself - Participant B keeps 20 ECU.

In the East situation, Participant B has the choice between:

- transferring the 25 ECU to Participant C - Participant C receives 50 ECU
- transferring 5 ECU to Participant C - Participant C receives 10 ECU - and keeping 20 ECU for himself - Participant B keeps 20 ECU.

In the South situation, Participant B has the choice between:

- transferring the 25 ECU to Participant A - Participant A receives 50 ECU
- transferring 5 ECU to Participant A - Participant A receives 10 ECU - and keeping 20 ECU for himself - Participant B keeps 20 ECU.

If Participant A has not sent 25 ECU to Participant B, Participant B does not make any decision.

Participant C receives an initial endowment of 30 ECU. Irrespective of the situation, he does not make any decision.

FIRST PART: STAGES

The first part of this session consists of four stages:

- Stage 1: All participants answer some questions.
- Stage 2: Participant A makes his decisions in the four situations.
- Stage 3: Participant B makes his decisions in the four situations.
- Stage 4: Participant A and Participant B answer some questions.

FIRST PART: COMPREHENSION QUESTIONNAIRE

Please complete the comprehension questionnaire that we will distribute to you. If you have any difficulty answering the questionnaire or when you have completed the questionnaire, raise your hand or press the red button on the side of your desk. We will answer your questions in private.

———————— End of the first set of instructions ————————

STAGE 1

In this stage, all participants answer some questions.

If you are a Participant B or a Participant C, you have to answer the following question: "Out of 3 Participants A randomly selected in today's session, how many of these Participants A will send 25 ECU to Participant B?". You have to answer this question for each situation: North, West, East and South.

If you are a Participant A or a Participant C, you have to answer the following question: "Out of 3 Participants B randomly selected in today's session, if Participant A has sent 25 ECU, how many of these Participants B will transfer the 25 ECU to another participant?". You have to answer this question for each situation: North, West, East and South.

How do the answers to these questions affect your earnings?

At the end of the session, for each role, one of the questions you answered during this stage will be randomly selected by the computer program. If your answer to this question is correct, you earn €1.

Example: Suppose you are Participant C and the randomly selected question is: "In the West situation, out of 3 Participants A randomly selected in today's session, how many of these Participants A will send 25 ECU to Participant B?". The computer program randomly select 3 Participants A among the Participants A in this session. If, in the West situation, "x" Participant(s) A among the 3 Participants A randomly selected has/have decided to send 25 ECU to Participant B, then, your answer is correct if you answered "x".

STAGE 2

In this stage, Participant A makes his decisions.

If you are Participant B or Participant C, you do not make any decision in this stage.

If you are Participant A, you decide whether or not to send 25 ECU to Participant B. You have to make this decision in each situation: North, West, East and South.

Which decision of Participant A determines the earnings of the group members?

At the end of the session, the computer program randomly selects the situation North, West, East or South. The decision made in the randomly selected situation determines the earnings of the group members. At the end of the session, all group members are informed of Participant A's randomly selected decision.

I you have any questions, raise your hand or press the red button on the side of your desk. We will answer your questions in private.

———————— End of the second set of instructions ————————

STAGE 3

In this stage, Participant B makes his decisions.

If you are Participant A or Participant C, you do not make any decision in this stage.

If you are Participant B, you decide how to distribute the 25 ECU you received between another participant (A or C) and yourself. You have to make this decision in each situation: North, West, East and South. Furthermore, in each situation, you have to make that decision for each possible prediction of Participant *A/C*.[26] To better understand, look at the screen example below. There are two pieces of information that appear in bold on the screen: information on the situation and information on the prediction of Participant *A/C*.

[26]Text between *... / ...* represents the two versions of the instructions. The first version corresponds to Treatment A and the second version corresponds to Treatment C.

figuresection

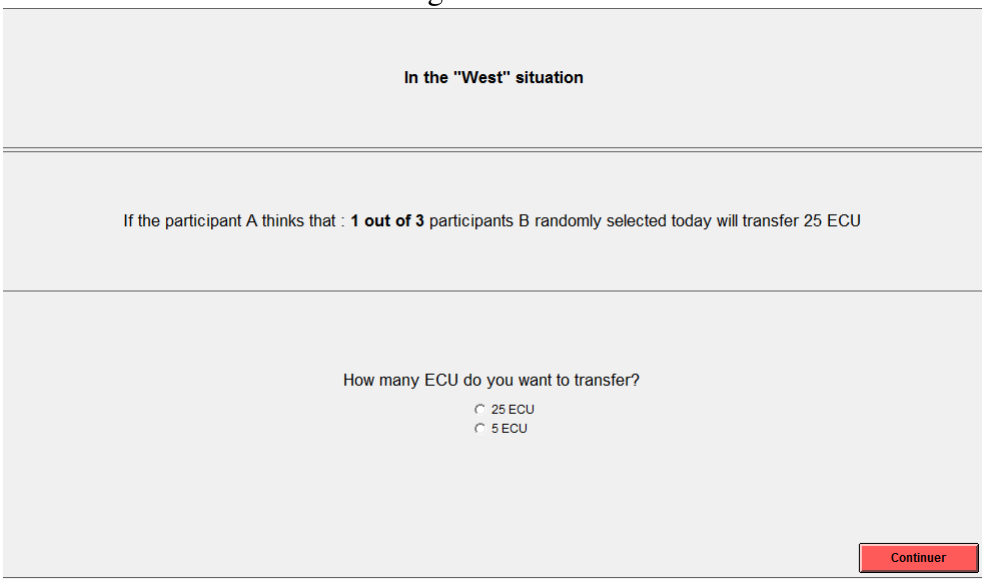


Figure 2.C.1: Screenshot in Treatment A

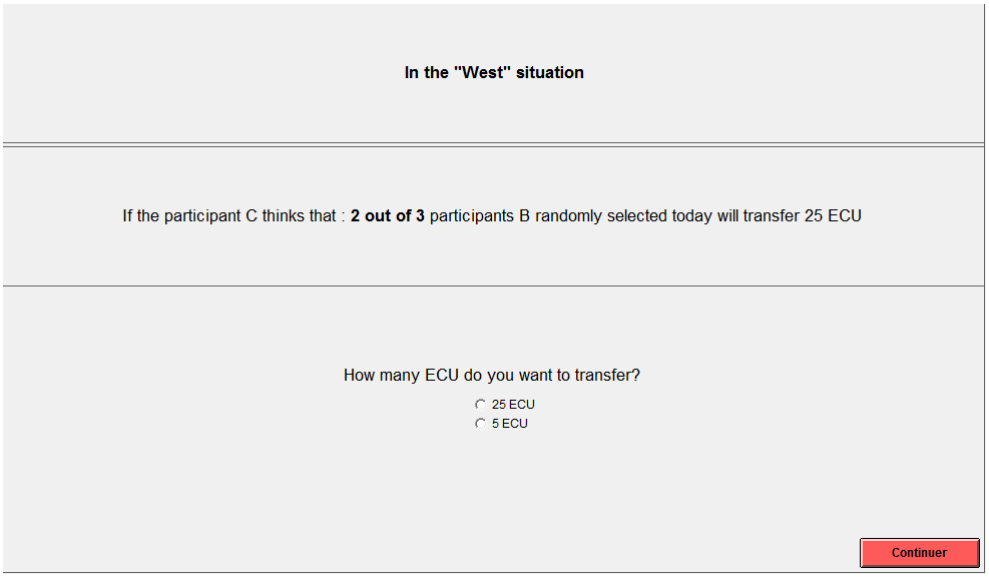


Figure 2.C.2: Screenshot in Treatment C

Information on the situation: You decide how to distribute the 25 ECU in each situation.

Example: In the screen above, you make your decision in the West situation.

Information on the prediction of Participant *A/C*: Remember that in stage 1, Participant *A/C* answered the following question for each situation: "Out of 3 Participants B randomly selected in today's session, if Participant A has sent 25 ECU, how many of these Participants B will transfer the 25 ECU to another participant?". There were four possible predictions: 0, 1, 2 or 3. You decide how to distribute the 25 ECU for each possible prediction of Participant *A/C*.

Example: In the screen above, you make your decision in the West situation, when Participant *A/C* in your group thinks that 2 out of 3 Participants B randomly selected today will transfer 25 ECU to another participant.

To summarize: You must therefore make 16 decisions:

- Four decisions corresponding to the four possible predictions of Participant *A/C* in the North situation
- Four decisions corresponding to the four possible predictions of Participant *A/C* in the West situation
- Four decisions corresponding to the four possible predictions of Participant *A/C* in the East situation
- Four decisions corresponding to the four possible predictions of Participant *A/C* in the South situation

However, only one of these decisions is susceptible to determine the earnings of the group members.

Which decision determines the earnings of the group members?

If Participant A has decided to send 0 EMU to Participant B, no decision of Participant B counts to determine the earnings of the group members.

If Participant A has decided to send 25 EMU to Participant B, a decision of Participant B determines the earnings of the group members. At the end of the session, the computer program randomly selects the situation North, West, East or South. Of the four decisions made by Participant B in the selected situation, the computer program then selects the decision that corresponds to the prediction that Participant *A/C* actually made in stage 1. This decision determines the earnings of the group members.

At the end of the session, all group members are informed of Participant B's randomly selected decision (if any).

Example: Suppose that the computer program randomly selects the West situation. Suppose that, to the question "In the West situation, out of 3 Participants B randomly selected in today's session, if Participant A has sent 25 ECU, how many of these B Participants will transfer the 25 ECU to another participant?", Participant *A/C* answered "2". Then, the computer program selects the decision that Participant B made when his screen displayed "West situation" and "Participant *A/C* thinks that 2 out of 3 Participants B randomly selected today will transfer 25 ECU to another participant " (see the example screen above). This decision determines the earnings of the group members.

If you have any questions, raise your hand or press the red button on the side of your desk. We will answer your questions in private.

———————— End of the third set of instructions ————————

STAGE 4

In this stage, Participant A and Participant B answer some questions.

If you are Participant C, you do not make any decisions in this stage. *If you are Participant A*, remember that, in stage 1, Participant B and Participant C answered the following question: "Out of 3 Participants A randomly selected in today's session, how many of these Participants A will send 25 ECU to Participant B?". They answered this question in each situation: North, West, East and South. You have to guess the answers of Participant B and of Participant C in your group.

If you are a Participant B, remember that, in stage 1, Participant A and Participant C answered the following question: "Out of 3 Participants B randomly selected in today's session, if Participant A has sent 25 ECU, how many of these Participants B will transfer the 25 ECU to another participant?". They answered this question in each situation: North, West, East and South. You have to guess the answers of Participant A and of Participant C in your group.

How do the answers to these questions affect your earnings?

At the end of the session, for each role, one of the questions you answered during this stage will be randomly selected by the computer program. If your answer to this question is correct, you earn €1.

Example: Suppose you are Participant A and the randomly selected question is: "According to Participant C in your group, in the situation West, among 3 Participants A randomly selected in today's session, how many of these Participants A will send 25 ECU to Participant B?". If, in stage 1, Participant C in your group answered that according to him, in the situation West, "x" Participant(s) A among the 3 Participants A randomly decided to send 25 EMU to Participant B, then, your answer is correct if you answered "x".

I you have any questions, raise your hand or press the red button on the side of your desk. We will answer your questions in private.

———————— End of the fourth set of instructions ————————

SECOND PART

In this part, the conversion rate is as follows: 10 EMU = €0.1.

There are fifteen periods. In each period, you have to choose the ECU allocation you prefer among nine allocations of ECU that will be proposed to you. An ECU allocation defines how many ECU you receive and how many ECU another participant X, randomly selected, receives.

Your earnings will be determined by one of your choices and by one of the choices of another participant Y, randomly selected. At the end of the session, a period will be randomly selected by the computer program, and the allocations chosen in this period determine your earnings:

- The allocation you have chosen during this period will be implemented for you and for another participant X, randomly selected.
- The allocation that another randomly selected participant Y has chosen during this period will be implemented for you and for him.

Your earnings in the second part are therefore the sum of your payoffs in these two selected allocations.

END OF THE SESSION

At the end of the session, you will be informed of the decisions that will have been selected at random to determine your payoffs (your decisions and those of other participants, if they affect your earnings) and of your final earnings.

Then, you will have to complete a final questionnaire.

At the end of the session, please remain seated and quiet until an experimentalist invites you to proceed to the payment room. Take your computer tag and your payment receipt with you. Leave the instructions on your desk.

If you have any questions, raise your hand or press the red button on the side of your desk. We will answer your questions in private.

———————— End of the last set of distributed instructions ————————

2.D Additional Results

2.D.1 Summary Statistics on Participants, by Treatment

Table 2.D.1: Summary Statistics on Participants, by Treatment

	Treatment A	Treatment C	Treatment Difference
% Women	61.22%	54.61%	No[2]
Mean age	21.90	22.42	No[1]
% Students	94.56%	93.62%	No[2]
% Business major	50.34%	54.61%	No[2]
Mean number of past participation in experiments	2.07	2.29	No[1]
Mean payoff (€)	17.09	17.03	No[1]
Number of sessions	8	9	
Number of subjects	147	141	

Notes: [1]Mann-Whitney rank-sum test; [2]Fisher exact test; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

2.D.2 Detailed Analysis of A-Subjects Behavior

We start by presenting in Table 2.D.2 descriptive statistics on the proportion of A-subjects who choose *In* given their first-order belief (α_{AB}) on the likelihood of *Right* choices by B-subjects.[27]

[27]Note that that across our 96 A-subjects, the majority of A-subjects (ranging from 45% to 65%) thought that none out of three possible B-subjects would choose *Right* after *In*, and only a small fraction (ranging from 3% to 9%) thought that the three out of three would choose *Right* after *In*.

Table 2.D.2: Proportion of *In* Choices Across Games, by First-Order Belief

% of *In* choices	Inv.	Rev-Inv.	Don.	Exp.
If $\alpha_{AB} = 0$	34.61%	69.76%	4.25%	80.95%
	(52)	(43)	(47)	(63)
If $\alpha_{AB} = 1/3$	48.00%	72.41%	17.85%	60.00%
	(25)	(29)	(28)	(20)
If $\alpha_{AB} = 2/3$	81.00 %	73.33%	64.70%	33.33%
	(16)	(15)	(17)	(6)
If $\alpha_{AB} = 1$	66.66%	66.66%	50.00%	100.00%
	(3)	(9)	(4)	(7)
All (96)	46.87%	70.83%	20.83%	75.00%

Notes: The sample size is in parentheses. α_{AB} is A's first-order belief that B chooses *Right* after *In*.

To estimate the impact of A's first-order beliefs and altruism sensitivity more formally, we report in Table 2.D.3 the results from Logit regressions on the probability that A-subjects choose *In* in each of the four games. For each game, there are two specifications. In the first specification, we regress A-subjects' *In* choice on α_{AB}, *i.e.*, their first-order belief on the likelihood of *Right* choices (to test 1), and on their SVO angle (to test 2). In line with our assumption of lexicographic altruistic preferences for A-subjects, we consider the SVO angle as a proxy of their altruism sensitivity. We control for the treatment and the order of the game. In the second specification, we add personality (self-reported risk aversion and patience) and socio-demographic controls (age, gender, number of previous experiments attended, and business major).

Regarding the influence of first-order beliefs (1), Table 2.D.3 reports average marginal effects estimated by random-effects Logit models. It shows that in the Investment and the Donation games the more A-subjects believed that *Right* was likely to be chosen by B-subjects, the more they chose *In*. These marginal effects are highly significant regardless of the specification. In contrast, in the Reversed-Investment and the Exploitation games, first-order beliefs do not significantly influence the frequency of *In* choices. This might be due to the majority of A-subjects being selfish, as it is usually found for trustees in comparable studies of trust games (see, e.g.,

Attanasi et al., 2013, 2019a). Recall that in both the Reversed-Investment and the Exploitation games, the predicted behavior of selfish A-subjects is *In* regardless of their α_{AB} (Table 2.3.1). This belief-independent behavior by the bulk of selfish A-subjects could prevent us from detecting the predicted positive correlation for non-selfish ones.

This analysis concludes that 1 is only supported for the Investment and the Donation games. R.A.1 states that as predicted, the frequency of *In* choices by A-subjects increases in their first-order belief about B-subjects choosing *Right* in the Investment and the Donation games, but not in the Reversed-Investment and the Exploitation games.

Table 2.D.3: Likelihood of A-Subjects Choosing *In*, by Games

	Investment		Rev-Investment		Donation		Exploitation	
FOB: α_{AB}	0.491***	0.397**	-0.062	-0.125	0.422***	0.463***	-0.103	-0.172
	(0.147)	(0.156)	(0.143)	(0.141)	(0.084)	(0.103)	(0.138)	(0.138)
SVO angle	0.011***	0.009***	0.006	0.005	0.007***	0.007***	-0.004	-0.004
	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)
Treatment A	0.160*	0.197**	0.095	0.069	-0.026	-0.003	-0.050	-0.020
	(0.089)	(0.090)	(0.094)	(0.093)	(0.071)	(0.070)	(0.086)	(0.086)
Order	0.011	0.036	-0.049	-0.042	0.006	0.024	0.085*	0.078
	(0.052)	(0.050)	(0.038)	(0.038)	(0.028)	(0.027)	(0.045)	(0.048)
Personality	No	Yes	No	Yes	No	Yes	No	Yes
Demographics	No	Yes	No	Yes	No	Yes	No	Yes
Observations	96	96	96	96	96	96	96	96
Log-likelihood	-54.361	-48.236	-56.075	-50.634	-33.255	-29.341	-50.963	-46.122
Prob>chi2	0.000	0.000	0.441	0.146	0.000	0.000	0.196	0.107
Pseudo R2	0.181	0.273	0.032	0.126	0.323	0.403	0.056	0.145

Notes: Table 2.D.3 reports the average marginal effects estimated by random-effects Logit models. Standard errors are in parentheses. “SVO angle” takes value between -7.8 and 45.9. “Order" is the rank order of the game, from 1 to 4. “Personality" controls correspond to the subjects’ self-reported risk aversion and patience. “Socio-Demographics" controls include age, gender, number of previous experiments attended, and business major. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Regarding the influence of altruism sensitivity on the frequency of *In* choices (2), Table 2.D.3 shows that in the Investment and the Donation games, the wider is the A-subject’s SVO angle, the higher is the likelihood to choose *In*. This effect is highly significant regardless of the specification. In the Reversed-Investment and the Exploitation games, it instead shows no significant effect of

A-subject's SVO angle over the likelihood to choose *In*. As for R.A.1, this absence of significance might be due to the over-representation of selfish A-subjects in our pool, who are predicted to choose *In* regardless of α_{AB}. However, recall that 2 is meant to show that the motivation behind the *In* choice of A is different across games. Therefore, finding a significant positive correlation between altruism sensitivity and trust in the two games where this was predicted and no correlation in the two games where a negative relation was predicted provides partial though solid support for 2.

R.A.2 states that in the Investment and the Donation games, where a positive choice-type correlation was predicted, there is a significant positive correlation between the frequency of *In* choices by A-subjects and their sensitivity to altruism, whereas in the Reversed-Investment and the Exploitation games, where a negative choice-type correlation was predicted, the correlation is not significant.

In addition, Table 2.D.3 shows that neither the treatment, nor the order of games had a significant effect on A-subjects' choices, with an exception in the Investment game for the treatment. The two previous results are robust to the inclusion of personality and socio-demographic controls.

In the spirit of the predictions of Table 2.3.1 on the altruism sensitivity ϕ_{Ah}, and of the separation among selfish, lightly-altruistic and highly-altruistic A-subjects on which 3 and 4 rely, we split A-subjects uniformly into these three categories according to their SVO angle. We define as "selfish" the A-subjects with a SVO angle in the interval $(Min, Median - 15\%)$ of the empirical distribution, as "lightly-altruistic" those with a SVO angle in the interval $(Median - 15\%, Median + 15\%)$, and as "highly-altruistic" those with a SVO angle in the interval $(Median + 15\%, Max)$.

To test 3, we consider the choices of the A-subjects who believe that *Left* is the most likely choice of B-subjects, *i.e.*, when A-subjects' $\alpha_{AB} \leq 1/3$. We also rely on the classification of the

subjects as selfish, lightly-altruistic and highly-altruistic. Focusing on selfish A-subjects, we find that the proportion of those who choose *In* is significantly lower in the Donation than in the Reversed-Investment game (2.70% vs 59.38%; proportion test, $p = 0.000$). Focusing on selfish and lightly-altruistic A-subjects, we find that this proportion is significantly lower in the Donation than in the Exploitation game (3.77% vs 80.77%; $p = 0.000$). Focusing on lightly-altruistic and highly-altruistic A-subjects, we find that this proportion is significantly lower in the Donation than in the Investment game (15.79% vs 48.89%; $p = 0.001$). These observations also hold if we release the constraints on the SVO angle of A-subjects, *i.e.*, if we consider the whole sample of A-subjects (proportion test, $p = 0.000$ for the three comparisons). For the Donation-Investment treatments comparison, this highlights the absence of unpredicted differences in the residual sub-samples of A-subjects, where the predicted choice is *Out* in both games (see Table 2.3.1). For the other two pairwise comparisons, the latter result indirectly confirms within our sample a negligible fraction of highly-altruistic subjects (*i.e.*, those for whom the predicted choice is *In* in the Donation game and *Out* in the other two games).

This analysis supports 3. R.A.3 states that, as predicted, for the A-subjects who believe that *Left* is the most likely action of B-subjects, the frequency of *In* choices in the Donation game is significantly lower than: (i) in the Reversed-Investment game for selfish types; (ii) in the Exploitation game for selfish and lightly-altruistic types; (iii) in the Investment game for lightly and highly-altruistic types.

Finally, to test 4, we consider the choices of the A-subjects who believe that *Right* is the most likely choice of B-subjects, *i.e.*, when A-subjects' $\alpha_{AB} > 2/3$. We find that the proportion of A-subjects who choose *In* in the Investment game is not significantly different than in the Reversed-Investment game regardless of the altruistic type (proportion test: $p = 1.000$ for selfish, $p = 0.898$ for lightly-altruistic, and $p = 0.177$ for highly-altruistic types). The proportion of selfish and lightly-altruistic A-subjects who choose *In* is higher in the Investment than in the Donation game (60% vs 50%; $p = 0.671$) but not significantly so. Finally, the proportion of

highly-altruistic A-subjects who choose *In* is significantly higher in the Investment than in the Exploitation game (100% vs 50%, $p = 0.021$). If we release the constraints on the SVO angle of A-subjects, *i.e.*, considering the whole sample of A-subjects, we still find that the proportion of A-subjects choosing *In* in the Investment game is not significantly different than in the Reversed-Investment game (78.94% vs. 70.83%, $p = 0.544$), and higher than in the Donation (61.90%, $p = 0.240$) and the Exploitation games (69.23%, $p = 0.533$).[28] This highlights the absence of unpredicted differences in the residual sub-samples of A-subjects (*i.e.*, with $\phi_{Ah} \in (0.30, 3.33]$), where the predicted choice is *In* in all games (see Table 2.3.1).

This analysis supports largely 4. R.A.4 states that, as predicted, for A-subjects who believe that Right is the most likely action of B-subjects, the frequency of In choices in the Investment game is: (i) not significantly different than in the Reversed-Investment game; (ii) higher, although not significantly so, than in the Donation game for lightly-altruistic and selfish types; (iii) significantly higher than in the Exploitation game for highly-altruistic types.

2.D.3 Within-Individual Analysis of B-Subjects' Patterns of Choices

We propose to classify the patterns of B-subjects' decisions into three main categories, depending on the consistency of their choices in the four games:

- consistent patterns when B-subjects always followed the same pattern of choices across the four games (52.08% of B-subjects);
- nearly consistent patterns when B-subjects followed the same pattern of choices in three games (20.83% of B-subjects);
- inconsistent patterns when B-subjects followed the same pattern of choices in at most two games (27.08% of B-subjects). The details of the choices of inconsistent subjects are available upon request.

[28] The last difference in proportion is not significant due to the small number of A-subjects with $\alpha_{Ah} \geq 2/3$ in the Exploitation game (only 13/96, see Table 2.D.2).

The left panel of Figure 2.D.1 displays the distribution of pattern categories for the B-subjects classified as guilt-averse in at least one game. The right panel of Figure 2.D.1 displays the same information for the B-subjects classified as selfish in at least one game. For both types of preferences, B-subjects who follow a consistent pattern of behavior in at least three games constitute the majority of our observations: 56.10% for guilt-averse subjects and 68,18% for selfish subjects.

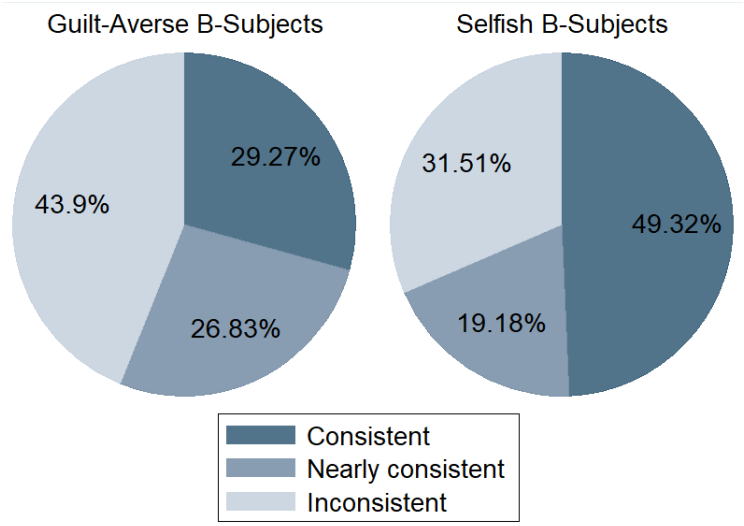


Figure 2.D.1: Distribution of B-Subjects Consistency of Behavior

Bibliography

Amdur, D. and Schmick, E. (2013). Does the direct-response method induce guilt aversion in a trust game? *Economics Bulletin*, 33(1):687–693.

Andreoni, J., Harbaugh, W. T., and Vesterlund, L. (2010). Altruism in experiments. In *Behavioural and experimental economics*, pages 6–13. Springer.

Andrighetto, G., Grieco, D., and Tummolini, L. (2015). Perceived legitimacy of normative expectations motivates compliance with social norms when nobody is watching. *Frontiers in Psychology*, 6:1413.

Attanasi, G., Battigalli, P., and Manzoni, E. (2016). Incomplete-information models of guilt aversion in the trust game. *Management Science*, 62(3):648–667.

Attanasi, G., Battigalli, P., Manzoni, E., and Nagel, R. (2019a). Belief-dependent preferences and reputation: Experimental analysis of a repeated trust game. *Journal of Economic Behavior & Organization*, 167:341–360.

Attanasi, G., Battigalli, P., and Nagel, R. (2013). Disclosure of belief-dependent preferences in a trust game. Technical report, No. 506, IGIER (Innocenzo Gasparini Institute for Economic Research), Bocconi University.

Attanasi, G., Rimbaud, C., and Villeval, M. C. (2019b). Embezzlement and guilt aversion. *Journal of Economic Behavior & Organization*, 167:409–429.

Balafoutas, L. and Fornwagner, H. (2017). The limits of guilt. *Journal of the Economic Science Association*, 3(2):137–148.

Balafoutas, L. and Sutter, M. (2017). On the nature of guilt aversion: Insights from a new methodology in the dictator game. *Journal of Behavioral and Experimental Finance*, 13:9–15.

Battigalli, P., Charness, G., and Dufwenberg, M. (2013). Deception: The role of guilt. *Journal of Economic Behavior & Organization*, 93:227–232.

Battigalli, P. and Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, 97(2):170–176.

Beck, A., Kerschbamer, R., Qiu, J., and Sutter, M. (2013). Shaping beliefs in experimental markets for expert services: Guilt aversion and the impact of promises and money-burning options. *Games and Economic Behavior*, 81:145–164.

Bellemare, C., Kröger, S., and Van Soest, A. (2008). Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities. *Econometrica*, 76(4):815–839.

Bellemare, C., Sebald, A., and Strobel, M. (2011). Measuring the willingness to pay to avoid guilt: estimation using equilibrium and stated belief models. *Journal of Applied Econometrics*, 26(3):437–453.

Bellemare, C., Sebald, A., and Suetens, S. (2017). A note on testing guilt aversion. *Games and Economic Behavior*, 102:233–239.

Bellemare, C., Sebald, A., and Suetens, S. (2018). Heterogeneous guilt sensitivities and incentive effects. *Experimental Economics*, 21(2):316–336.

Bellemare, C., Sebald, A., and Suetens, S. (2019). Guilt aversion in economics and psychology. *Journal of Economic Psychology*, 73:52–59.

Bracht, J. and Regner, T. (2013). Moral emotions and partnership. *Journal of Economic Psychology*, 39:313–326.

Chang, L. J., Smith, A., Dufwenberg, M., and Sanfey, A. G. (2011). Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron*, 70(3):560–572.

Charness, G. and Dufwenberg, M. (2011). Participation. *American Economic Review*, 101(4):1211–37.

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

Ciccarone, G., Di Bartolomeo, G., and Papa, S. (2020). The rationale of in-group favoritism: An experimental test of three explanations. *Games and Economic Behavior*, 124:554–568.

Dhami, S., Wei, M., and al Nowaihi, A. (2017). Public goods games and psychological utility: Theory and evidence. *Journal of Economic Behavior & Organization*.

Di Bartolomeo, G., Dufwenberg, M., Papa, S., and Passarelli, F. (2019). Promises, expectations & causation. *Games and Economic Behavior*, 113:137–146.

Dufwenberg, M., Gächter, S., and Hennig-Schmidt, H. (2011). The framing of games and the psychology of play. *Games and Economic Behavior*, 73(2):459–478.

Ederer, F. and Stremitzer, A. (2017). Promises and expectations. *Games and Economic Behavior*, 106:161–178.

Ellingsen, T., Johannesson, M., Tjøtta, S., and Torsvik, G. (2010). Testing guilt aversion. *Games and Economic Behavior*, 68(1):95–107.

Engler, Y., Kerschbamer, R., and Page, L. (2018). Guilt averse or reciprocal? looking at behavioral motivations in the trust game. *Journal of the Economic Science Association*, 4(1):1–14.

Ghidoni, R. and Ploner, M. (2020). When do the expectations of others matter? experimental evidence on distributional justice and guilt aversion. *Theory and Decision*, pages 1–46.

Hauge, K. E. (2016). Generosity and guilt: The role of beliefs and moral standards of others. *Journal of Economic Psychology*, 54:35–43.

Inderst, R., Khalmetski, K., and Ockenfels, A. (2019). Sharing guilt: How better access to information may backfire. *Management Science*, 65(7):3322–3336.

Ismayilov, H. and Potters, J. (2016). Why do promises affect trustworthiness, or do they? *Experimental Economics*, 19(2):382–393.

Kawagoe, T. and Narita, Y. (2014). Guilt aversion revisited: An experimental test of a new model. *Journal of Economic Behavior & Organization*, 102:1–9.

Khalmetski, K. (2016). Testing guilt aversion with an exogenous shift in beliefs. *Games and Economic Behavior*, 97:110–119.

Khalmetski, K., Ockenfels, A., and Werner, P. (2015). Surprising gifts: Theory and laboratory evidence. *Journal of Economic Theory*, 159:163–208.

Morell, A. (2019). The short arm of guilt–an experiment on group identity and guilt aversion. *Journal of Economic Behavior & Organization*, 166:332–345.

Ockenfels, A. and Werner, P. (2014). Scale manipulation in dictator games. *Journal of Economic Behavior & Organization*, 97:138–142.

Peeters, R. and Vorsatz, M. (2021). Simple guilt and cooperation. *Journal of Economic Psychology*, 82:102347.

Pelligra, V. (2011). Empathy, guilt-aversion, and patterns of reciprocity. *Journal of Neuroscience, Psychology, and Economics*, 4(3):161.

Regner, T. and Harth, N. S. (2014). Testing belief-dependent models. *Jena Economic Research Papers*.

Reuben, E., Sapienza, P., and Zingales, L. (2009). Is mistrust self-fulfilling? *Economics Letters*, 104(2):89–91.

Vanberg, C. (2008). Why do people keep their promises? an experimental test of two explanations 1. *Econometrica*, 76(6):1467–1480.

Yu, H., Shen, B., Yin, Y., Blue, P. R., and Chang, L. J. (2015). Dissociating guilt-and inequity-aversion in cooperation and norm compliance. *Journal of Neuroscience*, 35(24):8973–8975.

Chapter 3

Guilt Aversion and Information Acquisition

This chapter is co-authored with Alice Soldà.

3.1 Introduction

A large body of evidence has showed that individuals often care about the welfare of others.[1] These pro-social individuals face a trade-off between their monetary and moral motives. Hence, they might be tempted to exploit the uncertainty in their decision environment in order to reduce the tension between the two motives. In a seminal paper, later replicated by Larson and Capra (2009) and Feiler (2014), Dana et al. (2007) exposed this trade-off and coined the term "illusory" preferences, which refers to other-regarding preferences that fade away when introducing uncertainty on the relationship between

[1] For instance, people donate positive amounts of money to others without any strategic incentives to do so (Forsythe et al., 1994) or prefer more equitable monetary allocations over selfish ones (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). People donate more when their recipient expects to receive more (Bellemare et al., 2018; Attanasi et al., 2019), or lie less often when others' can infer their degree of dishonesty (Dufwenberg and Dufwenberg, 2018)

actions and outcomes.[2] The authors found significantly less generous behavior when participants were unsure about the consequences of their choice on others' payoffs.[3] This phenomenon has been supported by subsequent research. It has been shown that, in order to behave more selfishly, individuals manipulated their beliefs about others' intentions (Di Tella et al., 2015; Andreoni and Sanchez, 2020), remained strategically ignorant about the consequences of their choices (Grossman and Van Der Weele, 2017; Kajackaite, 2015), and took advantage of the presence of risk or ambiguity on whether donations were actually implemented (Haisley and Weber, 2010; Exley, 2016; Garcia et al., 2020). In the field, studies found that individuals avoided information on the environmental consequences of their actions (d'Adda et al., 2018) and news that could raise their empathy about refugees (Freddi, 2019).

This growing body of evidence has focused on *outcome*-based preferences, *i.e.*, preferences over payoffs. Yet, pro-social behavior can also be shaped by *belief*-based preferences, *i.e.*, preferences over payoffs *and* beliefs. To illustrate how these different types of preferences work, let's take an example. Ann offers Bob to work on a project together. Ann holds private expectations about how much Bob should work on the project. If Bob is a pure payoff-maximizer, he maximizes his utility function by providing zero effort, regardless of his beliefs about Ann's expectations. In contrast, Bob may be sensitive to Ann's expectations. If this is the case, Bob's beliefs about Ann's expectations can affect his level of effort in two directions. If Bob is guilt-averse, he maximizes his utility function by not disappointing others' expectations (Battigalli and Dufwenberg, 2007). Hence, his effort will increase with his beliefs about Ann's expectations. If,

[2]Throughout this paper we refer to this term more loosely as preferences that fade away in the presence of uncertainty in the decision environment. In particular, we use the concept of "subjective" preferences introduced by Spiekermann and Weiss (2016) to describe preferences defined over the epistemic state of the world rather than the actual state of the world.

[3]Serra-Garcia and Szech (2019) showed that this demand for less information is sensitive to the cost of avoiding the information.

instead, Bob is intention-based reciprocal, he maximizes his utility function by repaying (un)kindness with (un)kindness (Dufwenberg and Kirchsteiger, 2004, 2019). Therefore, an increase in his beliefs about Ann's expectations reduces how kind he perceives Ann to be. Consequently, his effort will decrease with his beliefs about Ann's expectations.

Results on the potentially "illusory" nature of belief-dependent preferences are mixed. Experimental studies addressing this question typically focused on reciprocity in trust games where the first-mover was uncertain about the second-mover's responsibility in the final outcome. In such context, the second-mover could, in principle, take advantage of this uncertainty to choose a less pro-social action than they would have in the absence of uncertainty. Van der Weele et al. (2014) showed that second-movers did not react to the introduction of uncertainty. In contrast, Regner (2018) and Regner and Matthey (2017) reported more selfish choices from second-movers. Furthermore, Friedrichsen et al. (2020) showed that reciprocity was lower when the first-mover's intentions were hidden.

To the best of our knowledge, there is no study on the potential "illusory" nature of guilt-aversion. Two related studies lead to contrasted conclusions. Grubiak et al. (2019) revealed that a substantial fraction of participants did not exploit the uncertainty about the relationship between their effort and their partner's outcome to break their promises. We interpret this finding as tentative evidence that guilt-aversion is not sensitive to the presence of situational excuses. In contrast, Inderst et al. (2019) found that second-movers in a trust game were less guilt-averse when the responsibility over the first-mover's payoff was shared between the first- and second-movers. This challenges the robustness of guilt aversion since participants were prone to exploit situational excuses to avoid being

pro-social.

In this paper, we investigate whether decision-makers with belief-dependent preferences self-servingly bias their information acquisition strategy in order to minimize the tension between their monetary interest and their belief-dependent concern. To address this question, we adapt the model of endogenous information acquisition by Spiekermann and Weiss (2016) to two specific belief-dependent preferences: guilt-aversion and reciprocity. In our framework, the second-mover ("trustee") in a modified trust game is uncertain about the first-mover's ("trustor") expectations, and can acquire information to resolve this uncertainty. Trustees can choose between two information sources. Each source provides either an informative signal about the trustor's expectations, or a null signal. A specific source can produce one of two types of informative signals, namely Low (*i.e.,* trustor's expectations are low) or High (*i.e.,* trustor's expectations are high), but never both.

Our model relies on two key features. First, we allow preferences to be "subjective" (*i.e.,* to depend on what trustees *believe* about the trustor's expectations). This is crucial as it creates an opportunity for subjective trustees to bias their acquisition of information in order to manipulate their beliefs about others' expectations. For objective trustees (*i.e.,* whose preferences depend on the trustor's *actual* expectations), there is no such opportunity given that the actual expectations of other players cannot be changed by any information acquisition strategy.[4] Second, we impose a coarse-grained mapping of beliefs, that is a mapping where beliefs can correspond to only three states: knowing with certainty that others' expectations are Low, knowing with certainty that they are High or

[4]Other models based on different mechanisms can also predicts strategic information acquisition by assuming belief-dependent preference, such as the model of moral constraints by (Rabin, 1995), self-signaling theories (*e.g.*, (Grossman and Van Der Weele, 2017)) or a model relying on an aversion to harm others (Chen et al., 2020).

remaining uncertain. Hence, acquiring information from a Low source never reveals with certainty that the expectations are High. Given this feature, full information avoidance is never an optimal strategy for subjective trustees, regardless of their belief-dependent motive.

We demonstrate that trustees with objective belief-dependent preferences should acquire signals from both sources of information, regardless of their belief-dependent motive, so that they can best condition their decisions on others' actual expectations. In contrast, trustees with subjective belief-dependent preferences should acquire the signal that reduce the conflict between their monetary payoff and their belief-dependent concern. Hence, a guilt-averse trustee will only seek a Low signal (as guilt-averse trustees can keep more for themselves by holding the belief that the trustor's expectations are low). Symmetrically, a reciprocal trustee will only seek a High signal (as reciprocal trustees can keep more for themselves by holding the belief that the trustor's expectations are high).

We then examine the different information acquisition strategies of trustees in an online experiment. As in the theoretical framework, we keep trustees uncertain about the trustors' outside option. In order to identify belief-dependent trustees, we first manipulate the trustors' outside option, which can be either Low or High. This manipulation aims to generate an exogeneous variation in trustees' beliefs: the higher the trustors' outside option, the higher the trustees' beliefs about the trustors' expectations. We then elicit trustees' return decisions conditional on learning that the trustor's outside option is either Low or High. Trustees are then unexpectedly given the opportunity to acquire information about the trustor's actual outside option, which will determine their final transfer. Trustees can either (i) acquire a High signal and learn with 50% chance that the

trustor outside option is High if it is truly the case, or remain uninformed; (ii) acquire a Low signal and learn with 50% chance that the trustor outside option is Low if it is truly the case, or remain uninformed; (iii) acquire both a Low and a High signal. If trustees learn the trustor's outside option, the conditional return corresponding to the actual outside option is implemented. If trustees remain uninformed about the trustor's outside option, the average of both conditional returns is implemented. Based on their choice of information acquisition, we can assess whether trustees strategically seek signals that are congruent with their monetary incentives rather seeking signals that maximize their information on the actual expectations of the trustor.

Consistent with the literature on "illusory" preferences, we find that 60.47% of guilt-averse trustees (*i.e.*, trustees who indicated a lower return choice conditional on learning that the trustor's outside option is Low) chose to acquire a Low signal only.[5] This information acquisition strategy is in line with our theoretical predictions of subjective guilt-aversion. In addition, we find that trustees who have the most money to lose from learning about the actual expectations of the trustor, are also the ones who are the most likely to engage in self-serving information acquisition strategies.

This paper closely relates to Spiekermann and Weiss (2016), who investigated whether norm-based preferences were "illusory" by giving participants an opportunity to strategically seek and/or avoid information. Our paper departs from theirs in two major ways. First, we focus on belief-based preferences instead of normative choices, *i.e.* on information about social norms. Furthermore, we allow behavior to be influenced by expectations both negatively and positively. Therefore, we can capture a broader range

[5] Note that because only one trustee exhibit intention-based reciprocity preferences consistent with the assumptions of our information acquisition model, we restricted our analyses to guilt-averse and belief independent trustees only.

of behavior than Spiekermann and Weiss (2016).

Our findings contribute to the literature on belief-dependent preferences. A long-standing debate in this literature concerns the best way to let participants learn about others' expectations.[6] Unlike most previous studies where the uncertainty about others' expectation is automatically resolved at the time where the action is implemented, we allow participants to resolve (or not) this uncertainty in a self-serving manner. Furthermore, we contribute to a recent strand of papers calling attention to the limited circumstances in which guilt-aversion is observed. For instance, Morell (2019) suggested that guilt aversion was observed only when social distance is small and Balafoutas and Fornwagner (2017) showed that it occurred when expectations were reasonable. Our findings show that uncertainty about others' beliefs dramatically affects pro-social behavior, which suggest that the extent to which belief-dependent preferences can sustain pro-social behavior may be overstated in the existing literature. Finally, our experimental design allows us to compare the relative weight of reciprocal vs. guilt-averse preferences in a randomly selected sample of the MTurk population. Consistent with Attanasi et al. (2010), our results suggest that guilt-aversion is the predominant motive in interactions in which the decision-maker's choice set is determined by a first-mover's willingness to blindly trust him/her.

In addition, our findings contribute to the literature on strategic information acquisition. While there is extensive evidence that individuals can deliberately remain ignorant about the consequences of their actions (see Golman et al., 2017 for a review),[7] a small body

[6] Beliefs about others' expectations can be self-reported (*e.g.*, Charness and Dufwenberg, 2006), directly communicated (*e.g.*, Ellingsen et al., 2010), induced via hypothetical alternatives (*e.g.*, Khalmetski et al., 2015) or induced via social norms (*e.g.*, Balafoutas and Sutter, 2017).

[7] In particular, Xiao and Bicchieri, 2012 showed that information avoidance can be relevant in the context of empirical expectations. The authors found that, when there is a small cost to acquire information, dictators avoid information on injuctive or descriptive norms about giving.

of research has now showed that individuals can also actively seek information if, in expected terms, selfish justifications become more available by doing so. When the information acquisition choice was binary, Fong and Oberholzer-Gee (2011) found that dictators who chose to acquire information about why their recipient was “poor", used it as an excuse to reduce their donations. When information was acquired sequentially, individuals stopped collecting information earlier when they liked early returns.[8] We differentiate ourselves from these lines of research by focusing on situations in which individuals can discriminate between different sources of information. In this literature, individuals have been shown to choose uninformative advisers (Shalvi et al., 2019), select positive feedback about their performance (Solda et al., 2019) or collect information on the undeservingness of recipients (Spiekermann and Weiss, 2016). Our findings add to this emergent literature by showing that individuals can also strategically discriminate between information sources when information relates to others' expectations.

The remaining of the paper is organised as follows. We develop our theoretical model in Section 3.2. In Section 3.3, we present the experimental design used to address our research question. Next, we derive our experimental hypotheses in Section 3.4. In Section 3.5, we describe our empirical results. Finally, we conclude in Section 3.6.

3.2 Theoretical Model

We introduce a Quasi-Trust game with incomplete information. The first mover (“trustor") decides between two actions: *In* or *Out*. If the trustor chooses *In*, the second mover

[8] For instance, information about their health (Ditto and Lopez, 1992), others' behavior (Smith et al., 2017) or the harm they may do to others (Chen et al., 2020).

("trustee") receives an endowment E to allocate between himself and the trustor.[9] The trustee returns y (with $0 \leq y \leq E$) to the trustor and keeps $E - y$ to himself. If the trustor chooses *Out*, the game ends and each player receives an outside option. The trustee receives $O^{trustee}$ and the trustor receives $O_{\omega}^{trustor}$ which depends on the state of the world $\omega \in \{L(ow), H(igh)\}$, with $O_H^{trustor} > O_L^{trustor}$. The trustor knows the state of the world with certainty when making her decision. In contrast, the trustee does not know the state of the world when choosing y, but knows that both states are equally likely to occur, that is: $p = p(\omega = L) = p(\omega = H) = 0.5$. The structure of the game is summarized in Figure 3.2.1.

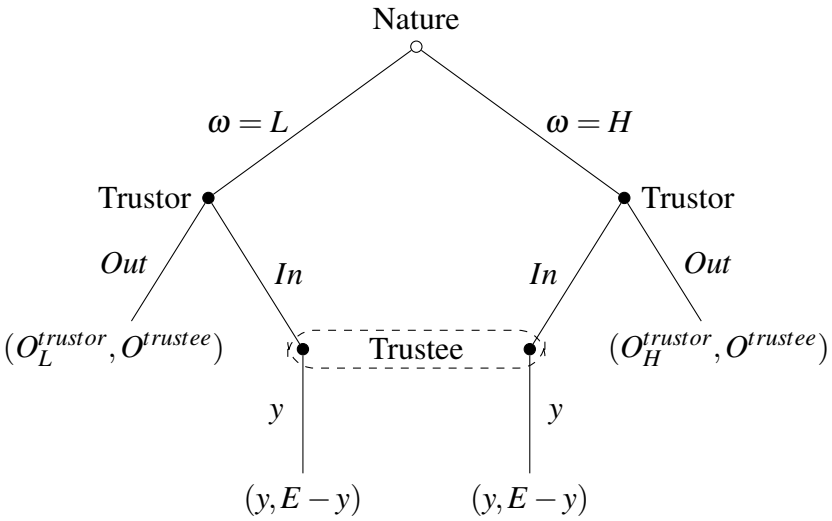


Figure 3.2.1: Trust game with High or Low outside option

We define $\phi_{\omega} \in [0, E]$, the trustor's first-order beliefs about his own payoff conditional on choosing *In*: $\phi_{\omega} = \mathbb{E}^{trustor}[y|In, \omega]$; and $\Phi_{\omega} \in [0, E]$, the trustee's second-order beliefs about the trustor's payoff conditional on choosing *In*: $\Phi_{\omega} = \mathbb{E}^{trustee}[\phi_{\omega}]$.

[9] For the sake of clarity, we use "she/her" when referring to the trustor and "he/him" when referring to the trustee.

3.2.1 Belief Formation

A payoff-maximizing trustor will choose *In* only if she expects to receive more from doing so than from choosing *Out*, i.e., when Equation 3.1 is satisfied.

$$\phi_\omega \geq O_\omega^{trustor} \tag{3.1}$$

Consequently, the higher the trustor's outside option is, the higher expectations of returns she signals by choosing *In* (i.e., the higher her first-order beliefs about her own payoff from choosing *In*): $\phi_L \leq \phi_H$. Inferring this, the trustee's second-order beliefs also increase in the trustor's outside option: $\Phi_L \leq \Phi_H$. This corresponds to an instance of psychological forward induction reasoning (Dufwenberg, 2002).[10] It leads to the following assumption.

Assumption. *Conditional on choosing* In*, trustors' first-order beliefs and trustees' second-order beliefs are higher when the outside option is High rather than Low.*

3.2.2 Belief-dependent preferences

We focus on the two main belief-dependent preferences that may be at play in a trust game: guilt-aversion and reciprocity. A guilt-averse trustee dislikes to disappoint others' expectations (Battigalli and Dufwenberg, 2007). His utility function corresponds to his material payoff minus the guilt he experiences (Equation 3.2). His guilt corresponds to the difference (if positive) between the trustor's expected payoff and the trustor's actual payoff. This difference is weighted by his sensitivity to guilt, denoted $\gamma_i \geq 0$

[10] Experimental evidence in favor of the psychological forward induction reasoning was provided by Woods and Servátka (2016).

(Equation 3.3).

$$u_{g,i}(y,\Phi_\omega) = (E-y) - g(y,\Phi_\omega) \tag{3.2}$$

$$\text{with } g_i(y,\Phi_\omega) = \gamma_i \cdot \max\{0,\,(\Phi_\omega - y)\} \tag{3.3}$$

If $\gamma_i < 1$, $u_{g,i}(y,\Phi_\omega)$ is maximized for $y^* = 0$ whereas, if $\gamma_i > 1$, $u_{g,i}(y,\Phi_\omega)$ is maximized for $y^* = \Phi_\omega$.[11] Proposition 1 follows below.

Proposition 1. *If $\gamma_i < 1$, guilt-averse trustees return $y^* = 0$. If $\gamma_i > 1$, guilt-averse trustees return $y^* = \Phi_\omega$.*

A reciprocal trustee likes to repay (un)kindness with (un)kindness (Dufwenberg and Kirchsteiger, 2004). His utility function corresponds to his material payoff plus the pleasure from reciprocating (Equation 3.4). The pleasure from reciprocating (Equation 3.5) is the product of the trustor's perceived kindness toward the trustee (Equation 3.6) and the trustee's kindness toward the trustor (Equation 3.7). The product is weighted by his sensitivity to reciprocity denoted $\rho_i \geq 0$. To define the trustor's perceived kindness, we compare the payoff that the trustee would receive, given the trustor's expectations ($E - \Phi_\omega$), to an equitable payoff which corresponds to the average between the worst and the best payoff he can receive, given the trustor's expectations ($\pi^e_{trustee}$). To define the trustee's kindness, we compare the payoff the trustor receives from the trustee's action (y) to an equitable payoff ($\pi^e_{trustor}$) which corresponds to the average between the worst

[11] If $\gamma_i = 1$, then the solution is indeterminate: $y^* \in [0, E]$.

and the best payoff she can receive from the trustee's action.

$$
\begin{aligned}
u_{r,i}(y,\Phi_\omega) &= (E-y)+r(y,\Phi_\omega) && (3.4)\\
\text{with } r_i(y,\Phi_\omega) &= \rho_i\cdot\lambda(\Phi_\omega)\cdot k(y) && (3.5)\\
\text{with } \lambda(\Phi_\omega) &= (E-\Phi_\omega)-\pi^e_{trustee}=(E-\Phi_\omega)-\frac{(E-\Phi_\omega)+O^{trustee}}{2}\\
&= \frac{E-O^{trustee}-\Phi_\omega}{2} && (3.6)\\
\text{with } k(y) &= y-\pi^e_{trustor}=y-\frac{0+E}{2}=y-\frac{E}{2} && (3.7)
\end{aligned}
$$

If $\Phi_\omega \geq E-O^{trustee}$, $u_{r,i}(y,\Phi_\omega)$ is maximized for $y^*=0$;[12] the trustor is perceived unkind, and the trustee responds with the least kind option possible. If $\Phi_\omega < E-O^{trustee}$, there is a trade-off between the reciprocity motive (responding with kindness) and the self-interest motive. Therefore, y^* depends on the relative weight of the material payoff, 1, and of the kindness function, $\rho_i\cdot\lambda(\Phi_\omega)$. If $\rho_i < \frac{2}{E-O^{trustee}-\Phi_\omega}$, $u_{r,i}(y,\Phi_\omega)$ is maximized for $y^*=0$, whereas if $\rho_i > \frac{2}{E-O^{trustee}-\Phi_\omega}$, $u_{r,i}(y,\Phi_\omega)$ is maximized for $y^*=E$.[13]

Proposition 2. *If $\Phi_\omega \geq E-O^{trustee}$, that is if the trustor is perceived as unkind, reciprocal trustees return $y^*=0$. If $\Phi_\omega < E-O^{trustee}$, that is the trustor is perceived as kind, and if (i) $\rho_i < \frac{2}{E-O^{trustee}-\Phi_\omega}$, reciprocal trustees return $y^*=0$ or (ii) $\rho_i > \frac{2}{E-O^{trustee}-\Phi_\omega}$, reciprocal trustees return $y^*=E$.*

For the remaining of the analysis, we introduce $f_i(y,\Phi)$, the psychological component of the utility function. When a trustee is guilt averse, then $f_i(y,\Phi)=g_i(y,\Phi)$; when a trustee is reciprocal, then $f_i(y,\Phi)=r_i(y,\Phi)$; and when the trustee does not have belief-dependent preferences, then $f_i(y,\Phi)=0$.

[12] When $\Phi_\omega = E-O^{trustee}$, this is a special case where the trustee is purely selfish because Equation 3.5 is null.

[13] If $\rho_i = \frac{2}{E-O^{trustee}-\Phi_\omega}$, then the solution is indeterminate: $y^*\in[0,E]$.

To reflect many real-life situations, we focus on cases where there exists a trade-off between the monetary payoff and the psychological component of the utility function, that is when $y^* > 0$ in at least one state of the world. Hence, we restrict our analysis to sufficiently guilt-averse ($\gamma_i > 1$) and sufficiently reciprocal ($\rho_i > \frac{2}{E-O^{trustee}-\Phi_\omega}$) trustees, and constrain our framework such that the trustor is always perceived as kind when choosing *In* ($\Phi_\omega < E - O^{trustee}$).

3.2.3 Information Acquisition Strategy

First, we define the trustees' decision problem under certainty about the true state of the world. Let $\hat{u}_{g,i}(\Phi_\omega) = \max_y u_{g,i}(y, \Phi_\omega)$ be a guilt-averse trustee's maximum utility achievable for a given expectation. This function decreases with Φ: the higher the expectations of the trustor, the less the trustee keeps for himself. Symmetrically, let $\hat{u}_{r,i}(\Phi_\omega) = \max_y u_{r,i}(y, \Phi_\omega)$ be a reciprocal trustee's maximum utility achievable for a given expectation. This function increases with Φ: the higher the expectations of the trustor, the lower the perceived kindness, the more the trustee keeps for himself.[14] The proof is provided in Section 3.A.1. Recalling our auxiliary assumption which states that $0 < \Phi_L < \Phi_H < E$, it follows that $\hat{u}_{g,i}(\Phi_L) > \hat{u}_{g,i}(\Phi_H)$ for a guilt-averse trustee and $\hat{u}_{r,i}(\Phi_L) < \hat{u}_{r,i}(\Phi_H)$ for a reciprocal trustee.

Second, we define the decision problem under uncertainty about the true state of the world in a situation where, before choosing how much to return, the trustee can acquire costless signals about the true state of the world. We define p' the updated probability of p after the acquisition of signal(s). The trustee can acquire one or two types of signals, represented by the random variables S_L and S_H. With probability s, the signal S_ω reveals that the state is ω given that the state is indeed ω ($S_\omega = \omega$), and with probability $1 - s$,

[14] For a reciprocal trustee we focus on the case where $y < \frac{E}{2}$. A detailed discussion of this choice is provided in Section 3.A.1.

the signal does not reveal the state of the world (null signal, $S_\omega = 0$). After a null signal, the trustee updates the probability p using Bayes' rule, which yields $p' = \frac{(1-s)p}{(1-s)p+(1-p)}$ after $S_L = 0$ and $p' = \frac{p}{p+(1-s)(1-p)}$ after $S_H = 0$. Receiving a null signal never removes the uncertainty. In fact, for any updated probability p', it exists $\varepsilon \geq 0$, the margin of tolerance "near certainty", such that $\varepsilon < p' < 1-\varepsilon$. Finally, if the trustee receives both $S_L = 0$ and $S_H = 0$, no update is necessary as the two signals cancel out each other: $p' = p$.

We distinguish between objective and subjective belief-dependent preferences. For an *objective* trustee, the psychological component depends on the true state of the world. Therefore, Φ_ω can take two values either $\Phi_\omega = \Phi_L$ in the Low state or $\Phi_\omega = \Phi_H$ in the High state. Under uncertainty, an objective trustee cannot be sure to choose the action that minimizes his guilt, or maximize his pleasure from reciprocity. Therefore, an objective guilt-averse (reciprocal) trustee must minimize (maximize) the expected psychological component given by: $p \cdot f_i(y, \Phi_L) + (1-p) \cdot f_i(y, \Phi_H)$.

For a *subjective* trustee, the psychological component depends on the epistemic state of the world. We follow Spiekermann and Weiss (2016) in proposing a coarse mapping of beliefs.[15] We define Φ_p the step function (Equation 3.8) corresponding to the three epistemic states: knowing that the true state is Low, $\Phi_{p'} = \Phi_L$, knowing that the true

[15] In the context of compliance to social norms, Spiekermann and Weiss (2016, p. 174) argue that "since degrees of beliefs are not observable in detail, it is unlikely that social norms take them as argument with any great precision. [...] This is mirrored in our everyday language rergarding normative choices, in which we rarely refer to degree of beliefs". We consider that the same reasoning applies for belief-dependent preferences.

state is High, $\Phi_{p'} = \Phi_H$, or not knowing the true state, $\Phi_{p'} = \Phi_U$.

$$\Phi_{p'} = \begin{cases} \Phi_L \text{ if } p' \geq 1-\varepsilon \\ \Phi_U \text{ if } \varepsilon < p' < 1-\varepsilon \\ \Phi_H \text{ if } p' \leq \varepsilon \end{cases} \tag{3.8}$$

What is the optimal information acquisition for a belief-dependent trustee depending on whether his preferences are objective or subjective? Recall that trustees can choose to acquire (i) S_L only, (ii) S_H only, or (iii) both S_L and S_H. Given y^*, an objective belief-dependent trustee maximizes his utility when his return matches the true state of the world, that is, when he knows the true state of the world. Hence, the information acquisition strategy that maximizes his utility is to acquire both signals, as it maximizes his chances to learn about the true state of the world. The proof is provided in Section 3.A.2.

Proposition 3. *Objective belief-dependent trustees acquire both signals, regardless of their belief-dependent motives.*

In contrast, a subjective belief-dependent trustee maximizes his utility if his return matches what *he believes* about the true state of the world. A subjective guilt-averse trustee minimizes the conflict between monetary payoff and guilt when he holds the belief that the state of the world is low (recall that under certainty, $\hat{u}_{g,i}(\Phi_L) > \hat{u}_{g,i}(\Phi_H)$). Consequently, a subjective guilt-averse trustee will sample information from the signal S_L only, which provides either information congruent with that beliefs, or no information. Symmetrically, a subjective reciprocal trustee minimize the conflict between monetary payoff and the pleasure from reciprocity when the state of the world is High (recall that under certainty, $\hat{u}_{r,i}(\Phi_H) > \hat{u}_{r,i}(\Phi_L)$). Consequently, a subjective reciprocal trustee will sample information from the signal S_H only, which provides either information congruent with that beliefs, or no information. The proof is provided in Section 3.A.3.

Proposition 4. *Subjective belief-dependent trustees with a coarse mapping of beliefs will acquire a Low signal only if they are guilt-averse, and acquire a High signal only if they are reciprocal.*

3.3 Design

In order to test our main theoretical predictions, we designed an experiment based on the modified trust game described in Section 3.2. Within this framework, we first introduced uncertainty about the trustors' expectations and then provided trustees' with an opportunity to acquire information to alleviate this uncertainty.

Trust game. At the beginning of the experiment, participants were randomly allocated to either the role of trustor or the role of trustee. Trustors faced two options: *Out* and *In*. If they chose *Out*, the game ended and both type of players received an outside option. The trustees' outside option was equal to 90 cents.[16] In contrast, the trustors' outside option depended on the game being played. If trustors chose *In*, they forewent their outside option. As a consequence, trustees received 200 cents to allocate between themselves and their matched trustor in increment of 15 cents. Players were informed of the entire payoff structure, including the existence of two equally likely outside options for trustors at the time of decision. Trustors were informed about their own outside option before they made their decision. In contrast, trustees did not know the trustors' actual outside option at the time of decision.

Outside option manipulation. In the Low game, trustors received 15 cents if they chose *Out*. In contrast, trustors received 75 cents if they chose *Out* in the High game. This feature of the design creates an exogenous variation in the participant's beliefs about

[16] All amounts are in USD.

the trustor's expected payoffs from choosing *In*. We operate under the assumption that trustors who chose *In* expect a return at least equal to the outside option that they were willing to forego. Therefore, conditional on choosing *In*, (i) trustors' first-order beliefs about their own payoff should be higher when the outside option is High rather than Low, and (ii) anticipating this, trustees' second-order beliefs about the trustors' payoff should also be higher when the outside option is High rather than Low.

Beliefs elicitation. Before both trustors and trustees made their decisions, we elicited their conditional beliefs about the trustors' expected payoffs from choosing *In*. To do so, we asked trustors to indicate how much they expected their trustee to send them if they chose *In*, both in the Low game and the High game. The elicitation of trustors' first-order beliefs was incentivized. Trustors' beliefs corresponding to the true state of the world were matched with their trustee's decision in the corresponding state of the world. If a trustor's belief was accurate, with a 15 cents margin of error, she was paid 50 cents. Symmetrically, we asked trustees to indicate how much trustors expected to receive if they chose *In*, both in the Low game and the High game. The elicitation of trustees' second-order beliefs was incentivized using the the same procedure described above. Trustees' beliefs corresponding to the true state of the world were matched with their trustor's belief in the corresponding state of the world. Trustees received 50 cents if their beliefs were accurate with a 15 cents margin of error.

Trustee's return choices. Trustees did not know their trustor's actual outside option at the time of decision. Hence, we elicited how much trustees wanted to send to the trustor both in case they *learned* that the outside option was Low (Decision Low) and in case they *learned* that the outside option was High (Decision High).[17] This key feature

[17]Note that because trustees did not observe the trustor's decision, they were asked to make a decision in the eventuality that their trustor chose *In* (strategy method).

of the design, inspired by the “menu" method of Bellemare et al. (2011), is crucial to identify trustees with belief-dependent preferences. Indeed, because trustors' outside options were designed to induce a shift in beliefs, eliciting trustees' returns conditional on their knowledge of the different outside options was equivalent to eliciting their choices conditional on the trustors' first-order beliefs.[18] Trustees were informed that if they learned before the end of the experiment that the trustor's actual outside option was Low, their Decision Low would be implemented. Symmetrically, if they learned that the trustor's actual option was High, their Decision High would be implemented. If they remained uninformed about their trustor's actual outside option, the trustor would receive the average of their Decision Low and their Decision High.[19]

Trustee's information acquisition. After making their conditional transfer decisions, trustees were unexpectedly offered the opportunity to acquire information about their trustor's outside option. To do so, we used the same procedure as in Spiekermann and Weiss (2016). Trustees faced two potential sources of information that took the form of four envelopes of two different colors: silver and gold. Trustees knew that if their trustor's outside option was Low, the information would be hidden in one of the two silver envelopes, and the three other envelopes would empty. If their trustor's outside option was High, the information would be hidden in one of the gold envelopes, and the three other envelopes would be empty. Trustees could open (i) a silver envelope, (ii) a gold envelope, or (iii) both a silver envelope and a gold envelope.[20] By opening a single envelope, trustees can strategically bias their information acquisition to learn only about one of the two signals. In contrast, opening both a silver and a gold envelope maximizes

[18] A similar procedure was used by Khalmetski (2016).

[19] Note that the model requires that $y_L < y_U < y_H$. Setting $y_U = \frac{y_L + y_H}{2}$ has the advantage to reflect the expected transfer under uncertainty without introducing probabilities to the participants.

[20] The order of presentation of the envelopes was randomized at the participant level.

trustees' chances to learn the trustor's actual outside option. The information acquisition procedure is summarized in Figure 3.3.1.

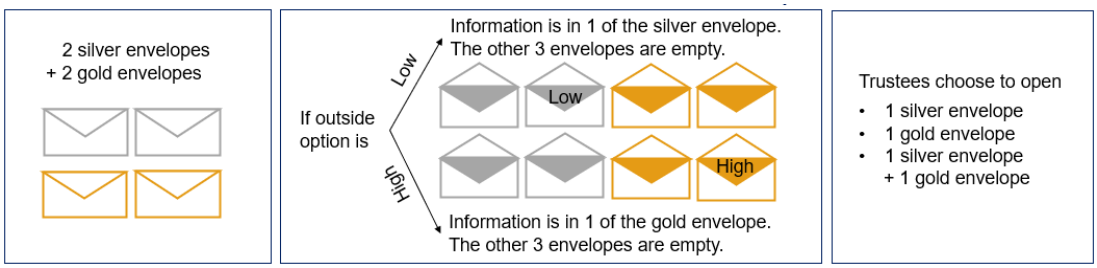


Figure 3.3.1: Choosing a source of information

Post-experimental questionnaires. By definition, participants' level of reasoning may affect their responsiveness to our treatment manipulation (as trustees needed to infer their trustor's expectations from their trustor's potential outside options). Hence, we elicited participants level of reasoning using a 2/3 beauty-contest game at the end of the experiment. Participants were asked to indicate a number between 0 and 100. Participants were rewarded with 100 cents if the number they indicated corresponded to two-third of the mean of the numbers indicated by all participants enrolled in the experiment. We also asked participants to report their age, gender, employment status, annual income and weekly expenditure. Finally, participants were asked to rate the clarity of the instructions using a scale from "extremely unclear" to "extremely clear". In addition, we asked trustees to explain their information acquisition decision in a free form format. They were rewarded 50 cents to provide an answer.

Procedures We conducted the experiment online on Amazon MTurk. We recruited a total of 320 participants from the United States of America.[21] Participation was restricted

[21] With that sample size, the minimum detectable effect size with statistical power at the recommended .80 level is 0.44 for comparisons of the proportion of each information acquisition strategy between belief-independent and belief-dependent trustees (Cohen, 2013), which is sufficient to detect an effect of half the magnitude of the one observed in Spiekermann and Weiss (2016).

to individuals over 18 years of age, who completed at least 300 HITs with an approval rate of at least 99%. Participants were randomly allocated the role of trustor or trustee at the beginning of the experiment. Pairs were formed after all participants had completed the experiment. During the experiment, participants could re-read the instructions at any time by clicking on a reminder button at the top of their screen.[22] Moreover, they had to answer a comprehension questionnaire correctly after the presentation of the instructions in order to proceed. Participants were paid less than 48 hours after the completion of the experiment.

3.4 Experimental Hypotheses

In Section 3.2, we assumed that a higher trustor's outside option increases both trustors' first-order beliefs and trustees' second-order belief about the trustor's payoff from choosing *In*. This is based on the rationale that a profit-maximizing trustor chooses *In* only if she expects to receive an amount at least equal to the outside option that she foregoes by doing so.[23]

Auxiliary Hypothesis 1. *Conditional on choosing* In*, trustors' first-order beliefs and trustees' second-order beliefs increase with the trustor's outside option.*

If Auxiliary Hypothesis 1 is verified, we can identify trustees' types based on their conditional transfer decisions: non belief-dependent (i.e., their return does not depend on the trustor's expectations), guilt-averse (i.e., their return increases with the trustor's

[22] The screens used in the experiment are provided in Section 3.B.

[23] Note that our manipulation of the outside option only implies that the distribution of "rational" beliefs conditional on choosing *In* has a higher *minimum* when the outside option is High rather than Low. Yet, it is possible that a trustor's first-order beliefs are constant across the two outside options but still within the range of "rational" beliefs.

expectations) and reciprocal (i.e., their return decreases with the trustor's expectations). Because we do not observe trustees guilt sensitivity γ_i or reciprocity sensitivity ρ_i, but only the choice y and the beliefs Φ_ω, we follow Dufwenberg et al. (2011) in identifying belief-dependent trustees approximately (rather than exactly). We contend that we can identify sufficiently guilt-averse players through the positive relationship between y and Φ_ω.[24] To identify reciprocal trustees, we adopt a similar method. We know that the higher Φ_ω, the less likely ρ_i is to exceed the threshold $\frac{2}{110-\Phi_\omega}$ above which it implies $y^* = 200$. Therefore, we can identify reciprocal trustees through the negative relationship between y and Φ_ω. This leads to our second auxiliary hypothesis.

Auxiliary Hypothesis 2. *The proportion of trustees identified as having belief-dependent preferences is strictly positive.*

Our main research question is to identify whether individuals who exhibit belief-dependent preferences bias their information acquisition strategy in a self-serving way. Belief-dependent preferences can be *objective*, that is, they can depend on the *actual* state of the world. If this is the case, participants should acquire both the Low signal and the High signal, regardless of their belief-dependent motive, as they maximize their utility when their return matches the actual state of the world. Hence, participants with objective belief-dependent preferences should open both a silver and a gold envelope, regardless of their belief-dependent motive. Yet, belief-dependent preferences can also be *subjective*, that is, they can depend on individuals' *beliefs* about the state of the world. In this case, our model predicts differentiated information acquisition strategies depending on the participant's belief-dependent motive. As described in Proposition 4, trustees with subjective belief-dependent preferences should acquire the signal that minimizes the tension between his monetary payoff and his belief-dependent concerns. Hence, a

[24] Remember that $y^* \in \{0; \Phi_\omega\}$.

guilt-averse trustee should only seek a Low signal. Symmetrically, a reciprocal trustee should only seek a High signal. This leads to our two main hypotheses.

Hypothesis 1. *Guilt-averse trustees are more likely to open only the silver envelope compared to belief-independent trustees.*

Hypothesis 2. *Reciprocal trustees are more likely to open only the gold envelope compared to belief-independent trustees.*

These hypotheses were pre-registered on AsPredicted.[25]

3.5 Experimental Results

In this section, we first evaluate whether beliefs are affected by the outside option manipulation. Then, we classify trustees' according to their type of preferences: belief-independent, guilt-averse or reciprocal. Finally, we assess how these preferences affect the trustees' information acquisition strategy.

3.5.1 Are beliefs affected by the outside option manipulation?

In this section, we assess whether Auxiliary Hypothesis 1 is verified, that is, whether trustors' first-order beliefs and trustees' second-order beliefs about trustors' payoff from choosing *In* are higher in High game than in the Low game.

[25] Link: https://aspredicted.org/blind.php?x=9md4uc.

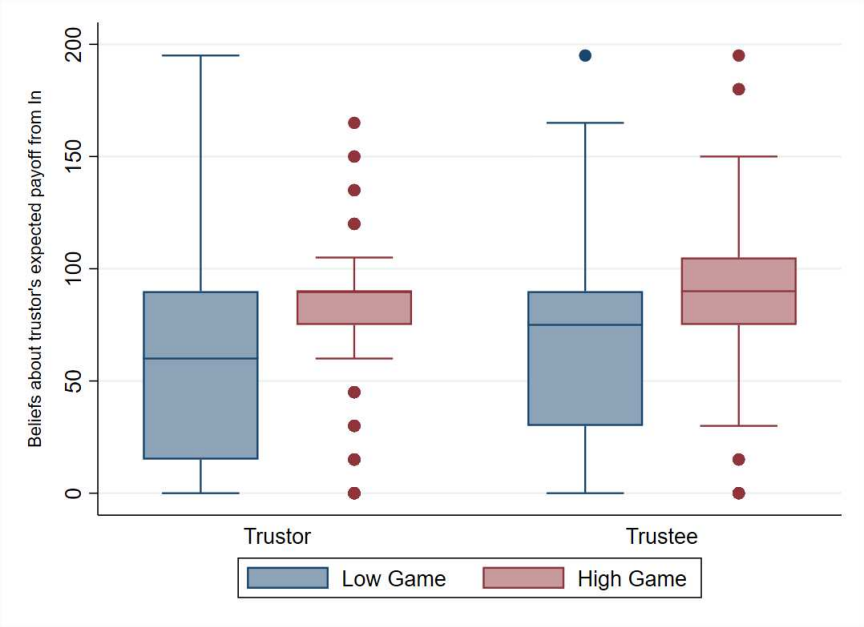
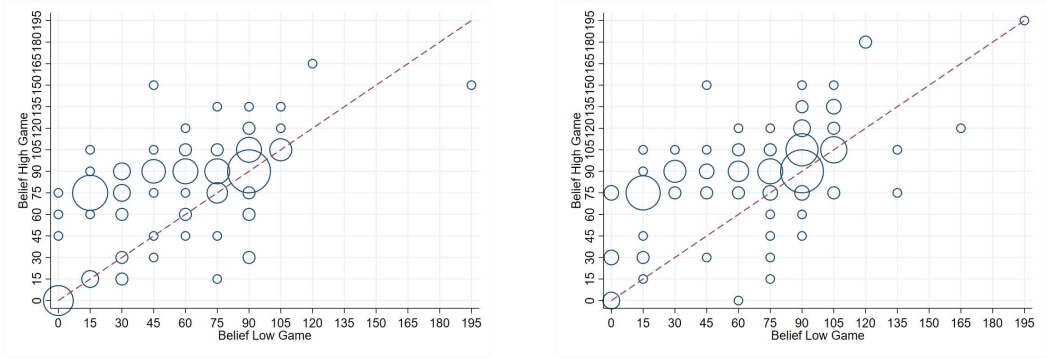


Figure 3.5.1: Distribution of trustors and trustees' beliefs about trustors' payoff from *In*.

At the aggregate level, Figure 3.5.1 shows that the trustors' median belief is lower in the Low game (median = 60 cents; interquartile range = 75)[26] than in the High game (med = 90 cents; iqr = 15). This difference is significant at the 0.01% level (Wilcoxon rank-sum test, $p < 0.001$).[27] Similarly, the trustees' median belief about trustors' belief is lower in the Low game (med = 75 cents; iqr = 60) than in the High game (med = 90 cents; iqr = 30). This difference is also significant at the 0.01% level (Wilcoxon signed rank test, $p < 0.001$). This yields Result 1 below.

Result 1. *Trustors' first-order beliefs and trustees' second-order beliefs are higher when the outside option is High rather than Low.*

[26] Respectively med and iqr, hereafter.
[27] All p-values are two-sided.



(a) Trustors' first-order beliefs

(b) Trustees' second-order beliefs

Figure 3.5.2: Distribution of individual beliefs about trustors' expected payoff from *In*

Turning to the individual level, Figure 3.5.2 displays the combination of beliefs about trustors' expected payoffs from choosing *In* in the Low game (x axis) and the High game (y axis). Beliefs of trustors are presented in the left panel and those of trustees in the right panel. Figure 3.5.2 shows that there is a lot of heterogeneity in responsiveness to the outside option manipulation. The majority of participants' beliefs verify Auxiliary Hypothesis 1. We find that 53.13% of trustors and 61.25% of trustees held higher beliefs in the High game than in the Low game (*i.e.*, observations above the 45 degree line). In contrast, 10% of trustors and 8.13% of trustees indicated higher beliefs in the Low game than in the high Game (*i.e.*, observations below the 45 degree line), while 28.75% of trustors and 38.75% of trustees indicated similar expectations regardless of the game being played (*i.e.*, observations on the 45 degree line). Interestingly, there seems to be a strong focal point around the egalitarian allocation with 50% of the participants with undifferentiated beliefs indicating beliefs at 90 cents in both games. To test our theoretical predictions, further analyses focus on the sub-sample of trustees who satisfied Auxiliary Hypothesis 1.

3.5.2 Are trustees motivated by belief-dependent preferences?

In this section, we classify trustees who satisfy Auxiliary Hypothesis 1 as guilt-averse, reciprocal or belief-independent based on their conditional transfers. Figure 3.5.3 displays the combinations of trustees' returns in the Low game (x axis) and the High game (y axis). Trustees who returned more in the High than in the Low game are classified as guilt-averse (*i.e.*, observations above the 45 degree line). In contrasts, trustees who returned more in the Low than in the High game are classified as reciprocal (*i.e.*, observations below the 45 degree line). Finally, trustees who returned the same amount regardless of the game are classified as belief-independent (*i.e.*, observations on the 45 degree line).

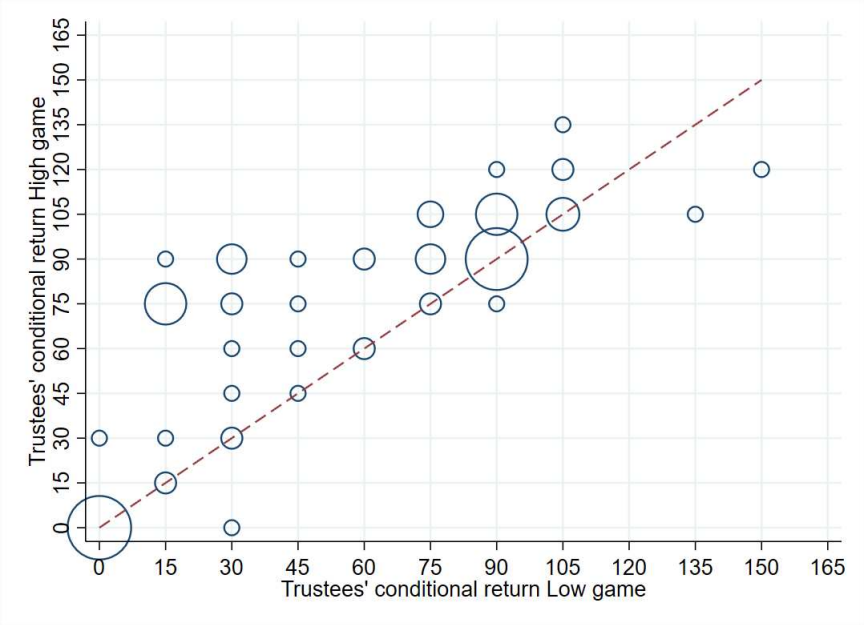


Figure 3.5.3: Trustees' return strategies

About half of the trustees can be classified as belief-independent (52.04%, n = 51), returning on average 50.00 cents (se = 6.14). This average return hides two focal points where the trustees' payoff is maximized (returning 0 cents) and where equality is maximized (returning 90 cents). We found that 43.88% (n = 43) of trustees can be

classified as guilt-averse. The average amount returned by guilt-averse trustees is 87.91 cents (se = 3.37) in the High game and 53.37 cents (se = 5.02) in the Low game. Only 4.08% of trustees can be classified as reciprocal (n = 4). The average amount returned by reciprocal trustees is 75 cents (se = 26.70) in the High game and 101.25 cents (se = 26.95) in the Low game (Wilcoxon signed rank test, $p = 0.058$). These observations yield our Result 2.

Result 2. *There is a positive proportion of belief-dependent trustees in our sample: 43.88% of trustees can be classified as guilt-averse, and 4.08% of trustees can be classified as reciprocal.*

3.5.3 How do belief-based preferences affect information acquisition?

We now examine whether trustees adopt different information acquisition strategies in belief-independent and belief-dependent trustees. As discussed in Section 3.2, we focus on trustees whose choices revealed a trade-off between their monetary and their belief-dependent motives. This excludes two reciprocal trustees who returned more than 90 (i.e. who cared only about their reciprocal motivation) and one additional reciprocal trustee who indicated that trustors' expected to receive more than 110 conditional on choosing *In* (*i.e.*, who perceived the trustor as unkind). Because this leaves us with only one reciprocal trustee that satisfies the criteria above, we restrict our analysis of trustees' information acquisition strategies to trustees that we classified as either belief-independent or guilt-averse.[28]

[28]Note that the unique reciprocal trustee with a tension between his/her material and reciprocal motivation chose to acquire a High signal only, which is consistent with subjective preferences.

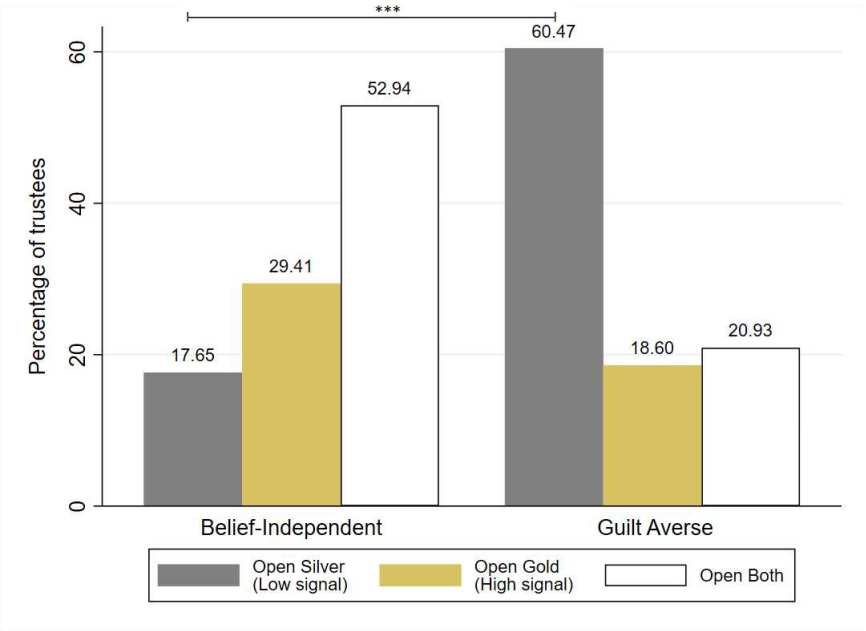


Figure 3.5.4: Distribution of information acquisition strategies for belief-independent and guilt-averse trustees.

Figure 3.5.4 displays the distribution of information acquisition strategies for belief-independent (left-hand side) and guilt-averse trustees (right-hand side). It shows that the majority of belief-independent trustees chose to open both envelopes (52.94%) and they did so significantly more than they would by chance (binomial test, $H0 = 0.33$, $p = 0.004$). We found that 17.65% of belief-independent trustees opened a silver envelope only, and 29.41% opened a gold envelope only. While our model makes no prediction on what belief-independent trustees should do, these results suggest that the default choice in the absence of strategic concerns, is to acquire as much information as possible.[29] The post-experimental questionnaire allows us to investigate potential explanation for the trustees' information acquisition strategy. It revealed that 75% of belief-independent trustees who chose to open both envelopes indicated that they did so out of curiosity.[30]

[29] Belief-independent trustees earn the same payoff irrespective of what they learn, as their conditional returns are the same regardless of the trustor's outside option.

[30] This result is based on the answers of the 43 out of 51 belief-independent trustees who did provide an answer. The distribution of answers can be found in Table 3.C.1 in Appendix.

In contrast to belief-independent trustees, the majority of guilt-averse trustees chose to open a silver envelop only (60.46%) and they did so significantly more than they would by chance (binomial test, $H0 = 0.33$, $p < 0.001$). In addition, the proportion of guilt-averse trustees who chose to open a silver envelope only is significantly higher than the proportion of belief-independent trustees who made the same choice (Pearson's chi-square test, $p < 0.001$) (Result 3).[31] These observations suggest that the majority of guilt-averse trustees exhibit an information acquisition strategy consistent with subjective preferences. In the post-experimental questionnaire, 79.17% of guilt-averse trustees who chose to open a silver envelope indicated that they did so because it maximized their payoffs.[32]

Result 3. *Guilt-averse trustees are more likely to open only the silver envelope compared to belief-independent trustees.*

Although our model was agnostic on the relative proportions of the different information acquisition strategies, it is noteworthy that 20.93% of guilt-averse trustees chose to open both envelopes and 18.60% opened a gold envelope only. Both of these proportions are significantly lower than the proportion of trustees opening a silver envelope only (Wilcoxon signed rank tests: $p = 0.004$ and $p = 0.002$, respectively). The trustees opening both envelopes displayed a behavior consistent with objective guilt-averse preferences. However, we contend that the proportion of trustees opening a gold envelope only corresponds to behavioral noise. Indeed, Figure 3.D.4 in Appendix shows that this share goes down to 5.56% when excluding trustees who reported that (i) they did not understand that their choice of envelopes was payoff-relevant (n=6) or (ii) the instructions

[31] These findings are consistent with the results from multinomial logit regressions reported in Table 3.C.2 in Appendix.

[32] This result is based on the answers of the 40 out of 43 guilt-averse trustees who did provide an answer.

were not “extremely clear" (n=24).

To further investigate the determinants of trustees' information acquisition strategy, we estimate a multinomial logit model in which the dependent variable is a categorical variable that summarizes the three information acquisition strategies. The main explanatory variable corresponds to the difference between the amount returned when the game is High and the amount returned when the game is Low. We control for participants' age and whether the participant indicated that she identified as a female, as well as participants' annual income and average weekly expenditures.

Table 3.5.1: Average marginal effects of monetary incentives on the likelihood of each sampling strategy

	Open Silver		Open Gold		Open both	
	(1)	(2)	(3)	(4)	(5)	(6)
Return(High) - Return(Low)	0.008***	0.009***	-0.001	-0.000	-0.007**	-0.009***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)
Ind. controls	No	Yes	No	Yes	No	Yes
Observations	94	94	94	94	94	94

Notes: Table 3.5.1 reports the average marginal effects of our multinomial logit model of the difference in conditional returns on the likelihood of a given sampling strategy. Controls include the amount guessed in the beauty contest game and socio-demographic characteristics (age, gender, annual income, weekly expenditure). Standard errors are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

The marginal effects are displayed in Table 3.5.1. We found that an increase in 10 cents in the difference in conditional returns increases the likelihood to open a silver envelope by up to 9 percentage points (columns (1) and (2)), and decreases the likelihood to open a gold envelope by up to 9 percentage points (columns (5) and (6)). These findings show that individuals who have the most money to lose from learning about a specific state of the world, are also the ones who are the most likely to engage in self-serving information

acquisition strategies.

Interestingly, female trustees were significantly more likely to open both envelopes ($AME = 0.244$, $p = 0.007$) and less likely to open the silver envelope only ($AME = -0.217$; $p = 0.025$), which suggest that female trustees are more likely to hold objective belief-dependent preferences than non-female trustees. This is supported by Table 3.C.3 in Appendix, which shows no differences in beliefs or conditional return decisions between male and female trustees (suggesting that this effect is not driven by differences in preference type between female and non-female trustees).

3.6 Discussion and Conclusion

Other-regarding preferences are prevalent in most human societies. However, the robustness of these preferences tends to be challenged in the presence of uncertainty in their decision environment. For instance, individuals with outcome-based preferences have been shown to exploit uncertainty about the relationship between actions and outcomes to behave more selfishly. In contrast, the literature on belief-dependent preferences has focused on situations where the uncertainty on others' expectations is automatically resolved when the action is implemented. Hence, one can wonder whether individuals with belief- based preferences would be prone to avoid the cost of following their moral conscience when they can manipulate the information they receive to resolve the uncertainty. In this paper, we investigated whether individuals with belief-dependent preferences select their source of information strategically in order to minimize the tension between their monetary interest and their belief-dependent motive.

We adapted the information acquisition model by Spiekermann and Weiss (2016) to study whether guilt-averse and reciprocal agents strategically acquire information about others'

expectations. Our model predicts that agents with objective belief-dependent preferences always prefer more information, while agents with subjective belief-dependent preferences strategically seek information that minimizes the tension between their monetary interest and their belief-dependent concern. We then tested our predictions in an online experiment. We designed a modified trust game in which we manipulate trustees' beliefs about trustors' expectations by varying trustors' outside option. We then elicited trustees' preferences by asking them to report their return choices conditionally on the trustors' outside option. Finally, trustees were given the opportunity to acquire information about the trustors' outside option.

We found that 60.47% of guilt-averse trustees chose to acquire only the signal that was congruent with their monetary incentives, consistent with our theoretical predictions for subjective preferences. Further analyses showed that individuals with the most differentiated return decisions were the most likely to engage in such self-serving information acquisition strategies. Finally, it is worth mentioning that a non-trivial fraction of our sample acquired information in a pattern consistent with objective belief-dependent preferences (20.93%).

Our main contribution is to show that a majority of individuals bias their information acquisition strategy towards self-serving signals to avoid the expected monetary cost of following their conscience (*i.e.*, implementing the return decision that corresponds to the true state of the world). In the literature on belief-dependent preferences, the uncertainty about others' expectations is typically resolved when actions are implemented (cf. menu method). Our experimental design deviates from this literature by allowing uncertainty about other's expectations either not to be resolved or to be resolved strategically, which is a more realistic information structure. Our findings suggest that previous research

have captured an upper bound of the positive impact of belief-dependent preferences on pro-social behavior.

Nonetheless, our paper has some limitations. Both our theoretical model and our experimental design consider the case of a coarse mapping of beliefs, *i.e.*, the belief about the state of the world is a step function. If we were to relax this feature and allow for linear beliefs, the optimal choice of a trustee with “illusory" belief-dependent preferences would be to avoid all information, a choice that was not possible in our experiment. However, this extension would require participants to be able to update their beliefs in a bayesian manner, which has been shown to be quite difficult (*e.g.*, Grether, 1980; Belot et al., 2012).

In terms of policy implications, it seems that nudging people towards pro-social outcomes can be done through their belief-dependent preferences only if others' expectations cannot be ignored. When others' expectations are uncertain and strategic information search is possible, a majority of belief-dependent individuals seek self-serving information, which eventually leads them to select the payoff-maximizing action without compromising their belief-dependent motives.

Bibliography

Andreoni, J. and Sanchez, A. (2020). Fooling myself or fooling observers? avoiding social pressures by manipulating perceptions of deservingness of others. *Economic Inquiry*, 58(1):12–33.

Attanasi, G., Battigalli, P., Nagel, R., et al. (2010). Disclosure of belief-dependent preferences in the trust game. In *BQGT*, pages 51–1.

Attanasi, G., Rimbaud, C., and Villeval, M. C. (2019). Embezzlement and guilt aversion. *Journal of Economic Behavior & Organization*, 167:409–429.

Balafoutas, L. and Fornwagner, H. (2017). The limits of guilt. *Journal of the Economic Science Association*, 3(2):137–148.

Balafoutas, L. and Sutter, M. (2017). On the nature of guilt aversion: Insights from a new methodology in the dictator game. *Journal of Behavioral and Experimental Finance*, 13:9–15.

Battigalli, P. and Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, 97(2):170–176.

Bellemare, C., Sebald, A., and Strobel, M. (2011). Measuring the willingness to pay to avoid guilt: estimation using equilibrium and stated belief models. *Journal of Applied Econometrics*, 26(3):437–453.

Bellemare, C., Sebald, A., and Suetens, S. (2018). Heterogeneous guilt sensitivities and incentive effects. *Experimental Economics*, 21(2):316–336.

Bolton, G. E. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American economic review*, 90(1):166–193.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6):1579–1601.

Chen, S., Heese, C., et al. (2020). Motivated information acquisition in social decisions. Technical report, University of Bonn and University of Mannheim, Germany.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

d'Adda, G., Gao, Y., Golman, R., and Tavoni, M. (2018). It's so hot in here: Information avoidance, moral wiggle room, and high air conditioning usage.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.

Di Tella, R., Perez-Truglia, R., Babino, A., and Sigman, M. (2015). Conveniently upset: Avoiding altruism by distorting beliefs about others' altruism. *American Economic Review*, 105(11):3416–42.

Ditto, P. H. and Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of personality and social psychology*, 63(4):568.

Dufwenberg, M. (2002). Marital investments, time consistency and emotions. *Journal of Economic Behavior & Organization*, 48(1):57–69.

Dufwenberg, M. and Dufwenberg, M. A. (2018). Lies in disguise–a theoretical analysis of cheating. *Journal of Economic Theory*, 175:248–264.

Dufwenberg, M., Gächter, S., and Hennig-Schmidt, H. (2011). The framing of games and the psychology of play. *Games and Economic Behavior*, 73(2):459–478.

Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and economic behavior*, 47(2):268–298.

Dufwenberg, M. and Kirchsteiger, G. (2019). Modelling kindness. *Journal of Economic Behavior & Organization*, 167:228–234.

Ellingsen, T., Johannesson, M., Tjøtta, S., and Torsvik, G. (2010). Testing guilt aversion. *Games and Economic Behavior*, 68(1):95–107.

Exley, C. L. (2016). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies*, 83(2):587–628.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3):817–868.

Feiler, L. (2014). Testing models of information avoidance with binary choice dictator games. *Journal of Economic Psychology*, 45:253–267.

Fong, C. M. and Oberholzer-Gee, F. (2011). Truth in giving: Experimental evidence on the welfare effects of informed giving to the poor. *Journal of Public Economics*, 95(5-6):436–444.

Forsythe, R., Horowitz, J. L., Savin, N. E., and Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic behavior*, 6(3):347–369.

Freddi, E. (2019). Do people avoid morally relevant information? evidence from the refugee crisis. *Review of Economics and Statistics*, pages 1–45.

Friedrichsen, J., Momsen, K., Piasenti, S., et al. (2020). Ignorance, intention and stochastic outcomes. Technical report.

Garcia, T., Massoni, S., and Villeval, M. C. (2020). Ambiguity and excuse-driven behavior in charitable giving. *European Economic Review*, 124:103412.

Golman, R., Hagmann, D., and Loewenstein, G. (2017). Information avoidance. *Journal of Economic Literature*, 55(1):96–135.

Grossman, Z. and Van Der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association*, 15(1):173–217.

Grubiak, K. et al. (2019). Exploring image motivation in promise keeping-an experimental investigation. Technical report, School of Economics, University of East Anglia, Norwich, UK.

Haisley, E. C. and Weber, R. A. (2010). Self-serving interpretations of ambiguity in other-regarding behavior. *Games and economic behavior*, 68(2):614–625.

Inderst, R., Khalmetski, K., and Ockenfels, A. (2019). Sharing guilt: How better access to information may backfire. *Management Science*, 65(7):3322–3336.

Kajackaite, A. (2015). If i close my eyes, nobody will get hurt: The effect of ignorance on performance in a real-effort experiment. *Journal of Economic Behavior & Organization*, 116:518–524.

Khalmetski, K. (2016). Testing guilt aversion with an exogenous shift in beliefs. *Games and Economic Behavior*, 97:110–119.

Khalmetski, K., Ockenfels, A., and Werner, P. (2015). Surprising gifts: Theory and laboratory evidence. *Journal of Economic Theory*, 159:163–208.

Larson, T. and Capra, C. M. (2009). Exploiting moral wiggle room: Illusory preference for fairness? a comment. *Judgment and decision Making*, 4(6):467.

Morell, A. (2019). The short arm of guilt–an experiment on group identity and guilt aversion. *Journal of Economic Behavior & Organization*, 166:332–345.

Rabin, M. (1995). Moral preferences, moral constraints, and self-serving biases.

Regner, T. (2018). Reciprocity under moral wiggle room: Is it a preference or a constraint? *Experimental Economics*, 21(4):779–792.

Regner, T. and Matthey, A. (2017). Actions and the self: I give, therefore i am? *Jena Economic Research Papers*, 2017:018.

Serra-Garcia, M. and Szech, N. (2019). The (in) elasticity of moral ignorance.

Shalvi, S., Soraperra, I., van der Weele, J. J., and Villeval, M.-C. (2019). Shooting the messenger? supply and demand in markets for willful ignorance. Technical report, Tinbergen Institute Discussion Paper.

Smith, M. K., Trivers, R., and von Hippel, W. (2017). Self-deception facilitates interpersonal persuasion. *Journal of Economic Psychology*, 63:93–101.

Solda, A., Ke, C., Page, L., and Von Hippel, W. (2019). Strategically delusional. *Experimental Economics*, pages 1–28.

Spiekermann, K. and Weiss, A. (2016). Objective and subjective compliance: A norm-based explanation of 'moral wiggle room'. *Games and Economic Behavior*, 96:170–183.

Van der Weele, J. J., Kulisa, J., Kosfeld, M., and Friebel, G. (2014). Resisting moral wiggle room: how robust is reciprocal behavior? *American economic Journal: microeconomics*, 6(3):256–64.

Woods, D. and Servátka, M. (2016). Testing psychological forward induction and the updating of beliefs in the lost wallet game. *Journal of Economic Psychology*, 56:116–125.

Xiao, E. and Bicchieri, C. (2012). Words or deeds? choosing what to know about others. *Synthese*, 187(1):49–63.

Appendix

3.A Proofs adapted from Spiekermann and Weiss (2016)

3.A.1 Proof on the variation of $\hat{u}$ with respect to Φ

We want to show that $\hat{u}'_{g,i}(\Phi_\omega) \leq 0$ for guilt-averse trustees and $\hat{u}'_{r,i}(\Phi_\omega) \geq 0$ for reciprocal trustees. According to the envelope theorem, the total derivative at point $y*$ is equal to the following partial derivative:

$$\begin{aligned} \hat{u}'_i(\Phi) &= \frac{\partial}{\partial \Phi} u_i(y, \Phi)\Big|_{y=y^*} \\ &= \frac{\partial}{\partial \Phi}[(E - y^*) - f_i(y^*, \Phi)] \end{aligned} \tag{3.9}$$

For a guilt-averse trustee, it yields the following. Note that, by construction $\Phi_\omega \geq y^*$, therefore $max\{0, (\Phi_\omega - y^*)\} = \Phi_\omega - y^*$.

$$\begin{aligned} \hat{u}'_{g,i}(\Phi_\omega) &= \frac{\partial}{\partial \Phi_\omega}[(E - y^*) - \gamma_i(\Phi_\omega - y^*)] \\ &= -\gamma_i \leq 0 \end{aligned} \tag{3.10}$$

For a reciprocal trustee, it yields the following:

$$\begin{aligned}\hat{u}'_{r,i}(\Phi_\omega) &= \frac{\partial}{\partial \Phi_\omega}[(E-y^*)+\rho_i(\frac{E-90-\Phi_\omega}{2})(y^*-\frac{E}{2})] \\ &= \underbrace{-\frac{\rho_i}{2}}_{\leq 0}\underbrace{(y^*-\frac{E}{2})}_{?} \end{aligned} \tag{3.11}$$

As highlighted in Section 3.2, we focus on the information acquisition of trustees who face a trade-off between their monetary and belief-dependent motives, which exclude reciprocal trustees who give more than half of their endowment to the trustor: $y > \frac{E}{2}$ (*i.e.*, trustees who only care about their belief-dependent motive). This restriction is consistent with the typical behavior observed in the literature: the meta-study by Johnson and Milsin (2011) on behavior in the trust-game shows that approximately 90% of trustees return less than half of their endowment. It is also consistent with the behavior observed in our experimental data, where 88.75% of trustees returned less than half the endowment in the Low game, and 81.25% in the High game. Restricting our analysis to the case where $y < \frac{E}{2}$, we can conclude that $\hat{u}'_{r,i}(\Phi) > 0$.

3.A.2 Proof of Proposition 3

In this proof to simplify the notation, we denote $E-y$ as $v(y)$.

When acquiring the signal is S_L, the expected utility of an objective guilt-averse trustee is the weighted sum of the trustee's guilt when the true state is $\omega = L$ and the trustee knows it (with probability ps), when the true state is S_L but the trustee does not know it (with probability $p(1-s)$), and when the true state is S_H (with probability $(1-p)$).

$$\begin{aligned} Eu_L &= ps \cdot u_{g,i}(y_L^*, \Phi_L) + p(1-s)\cdot v(y_U^*, \Phi_L) - p(1-s)\cdot g_i(y_U^*, \Phi_L) \\ &\quad +(1-p)\cdot v(y_U^*, \Phi_H) - (1-p)\cdot g_i(y_U^*, \Phi_H) \\ &= ps \cdot u_{g,i}(y_L^*, \Phi_L) + (1-ps)\cdot v(y_U^*) - p(1-s)\cdot g_i(y_U^*, \Phi_L) \\ &\quad -(1-p)\cdot g_i(y_U^*, \Phi_H) \end{aligned} \tag{3.12}$$

Similarly, when acquiring the signal is S_L, the expected utility of an objective reciprocal trustee is given by the following equation.

$$\begin{aligned} Eu_L &= ps \cdot u_{r,i}(y_L^*, \Phi_L) + (1-ps)\cdot v(y_U^*) + p(1-s)\cdot r_i(y_U^*, \Phi_L) \\ &\quad +(1-p)\cdot r_i(y_U^*, \Phi_H) \end{aligned} \tag{3.13}$$

When acquiring the signal is S_H, the expected utility of an objective guilt-averse trustee is the weighted sum of the trustee's guilt when the true state is $\omega = H$ and the trustee knows it (with probability $(1-p)s$), when the true state is S_H but the trustee does not know it (with probability

$(1-p)(1-s)$), and when the true state is S_L (with probability p).

$$\begin{aligned} Eu_H &= (1-p)s\cdot u_{g,i}(y_H^*,\Phi_H)+(1-p)(1-s)\cdot v(y_U^*,\Phi_H) \\ &\quad -(1-p)(1-s)\cdot g_i(y_U^*,\Phi_H)+p\cdot v(y_U^*,\Phi_L)-p\cdot g_i(y_U^*,\Phi_L) \\ &= (1-p)s\cdot u_{g,i}(y_H^*,\Phi_H)+(1-s+ps)\cdot v(y_U^*)-(1-p)(1-s)\cdot g_i(y_U^*,\Phi_H) \\ &\quad -p\cdot g_i(y_U^*,\Phi_L) \end{aligned} \tag{3.14}$$

Similarly, when acquiring the signal is S_H, the expected utility of an objective reciprocal trustee is given by the following equation.

$$\begin{aligned} Eu_H &= (1-p)s\cdot u_{r,i}(y_H^*,\Phi_H)+(1-s+ps)\cdot v(y_U^*)+(1-p)(1-s)\cdot r_i(y_U^*,\Phi_H) \\ &\quad +p\cdot r_i(y_U^*,\Phi_L) \end{aligned} \tag{3.15}$$

When acquiring both signals, the expected utility of an objective guilt-averse trustee is the weighted sum of the trustee's guilt when the true state is $\omega = L$ and the trustee knows it (with probability ps), when the true state is $\omega = H$ and the trustee knows it (with probability $(1-p)s$) when the true state is S_L but the trustee does not know it (with probability $p(1-s)$), and when the true state is S_H but the trustee does not know it (with probability $((1-p)(1-s))$.

$$\begin{aligned} Eu_{LH} &= ps\cdot u_{g,i}(y_L^*,\Phi_L)+(1-p)s\cdot u_{g,i}(y_H^*,\Phi_H)+p(1-s)\cdot v(y_U^*,\Phi_L)-p(1-s)\cdot g_i(y_U^*,\Phi_L) \\ &\quad +(1-p)(1-s)\cdot v(y_U^*,\Phi_H)-(1-p)(1-s)\cdot g_i(y_U^*,\Phi_H) \\ &= ps\cdot u_{g,i}(y_L^*,\Phi_L)+(1-p)s\cdot u_{g,i}(y_H^*,\Phi_H)+(1-s)\cdot v(y_U^*) \\ &\quad -p(1-s)\cdot g_i(y_U^*,\Phi_L)-(1-p)(1-s)\cdot g_i(y_U^*,\Phi_H) \end{aligned} \tag{3.16}$$

Similarly, when acquiring both signals, the expected utility of an objective reciprocal trustee is given by the following equation.

$$\begin{aligned} Eu_{LH} &= ps\cdot u_{r,i}(y_L^*,\Phi_L)+(1-p)s\cdot u_{r,i}(y_H^*,\Phi_H)+(1-s)\cdot v(y_U^*)+p(1-s)\cdot r_i(y_U^*,\Phi_L) \\ &\quad +(1-p)(1-s)\cdot r_i(y_U^*,\Phi_H) \end{aligned} \tag{3.17}$$

We compare the expected utilities of receiving signal S_H (Eu_H) to receiving both signals (Eu_{LH}).

$$\begin{aligned} Eu_{LH}-Eu_H &= ps\cdot u_{g,i}(y_L^*,\Phi_L)+(1-p)s\cdot u_{g,i}(y_H^*,\Phi_H)+(1-s)\cdot v(y_U^*)-p(1-s)\cdot g_i(y_U^*,\Phi_L) \\ &\quad -(1-p)(1-s)\cdot g_i(y_U^*,\Phi_H)-(1-p)s\cdot u_{g,i}(y_H^*,\Phi_H)-(1-s+ps)\cdot v(y_U^*) \\ &\quad +(1-p)(1-s)\cdot g_i(y_U^*,\Phi_H)+p\cdot g_i(y_U^*,\Phi_L) \\ &= ps\cdot[u_{g,i}(y_L^*,\Phi_L)-v(y_U^*)+g_i(y_U^*,\Phi_L)] \\ &= ps\cdot[\underbrace{u_{g,i}(y_L^*,\Phi_L)-u_{g,i}(y_U^*,\Phi_L)}_{>0}] \end{aligned} \tag{3.18}$$

Equation 3.18 is positive since, given Φ_L, utility is maximal at $\hat{u}_{g,i}(\Phi_L) = u_{g,i}(y^*_L, \Phi_L)$. Using the same reasoning, it yields to the following equation for subjective reciprocal trustees.

$$Eu_{LH} - Eu_H = ps \cdot \underbrace{[u_{r,i}(y^*_L, \Phi_L) - u_{r,i}(y^*_U, \Phi_L)]}_{>0} \tag{3.19}$$

We compare the expected utilities of receiving signal S_L (Eu_L) to receiving both signals (Eu_{LH}).

$$\begin{aligned} Eu_{LH} - Eu_L &= ps \cdot u_{g,i}(y^*_L, \Phi_L) + (1-p)s \cdot u_{g,i}(y^*_H, \Phi_H) + (1-s) \cdot v(y^*_U) - p(1-s) \cdot g_i(y^*_U, \Phi_L) \\ &\quad -(1-p)(1-s) \cdot g_i(y^*_U, \Phi_H) - ps \cdot u_{g,i}(y^*_L, \Phi_L) - (1-ps) \cdot v(y^*_U) \\ &\quad +p(1-s) \cdot g_i(y^*_U, \Phi_L) + (1-p) \cdot g_i(y^*_U, \Phi_H) \\ &= (1-p)s \cdot [u_{g,i}(y^*_H, \Phi_H) - u_{g,i}(y^*_U, \Phi_H)] > 0 \end{aligned} \tag{3.20}$$

Equation 3.20 is positive since, given Φ_H, utility is maximal at $\hat{u}_{g,i}(\Phi_H) = u_{g,i}(y^*_H, \Phi_H)$. Using the same reasoning, it yields to the following equation for subjective reciprocal trustees.

$$Eu_{LH} - Eu_L = (1-p)s \cdot [u_{r,i}(y^*_H, \Phi_H) - u_{r,i}(y^*_U, \Phi_H)] > 0 \tag{3.21}$$

We can conclude that taking both signals is the preferred choice for objective belief-dependent trustees.

3.A.3 Proof of Proposition 4

The expected utility of acquiring signal S_L for a subjective trustee corresponds to the weighted sum of the trustee's utility when the state is $\omega = L$ and the trustee knows it with probability (with probability ps), and when the trustee is uncertain about the state (with probability $1 - ps$).

$$Eu_L = ps \cdot \hat{u}_i(\Phi_L) + (1-ps) \cdot \hat{u}_i(\Phi_U) \tag{3.22}$$

Symmetrically, The expected utility of acquiring signal S_L for a subjective trustee corresponds to the weighted sum of the trustee's utility when the state is $\omega = L$ and the trustee knows it with probability (with probability $(1-p)s$), and when the trustee is uncertain about the state (with probability $1 - s + ps$).

$$Eu_H = (1-p)s \cdot \hat{u}_i(\Phi_H) + (1-s+ps) \cdot \hat{u}_i(\Phi_U) \tag{3.23}$$

Finally, the expected utility of acquiring signal S_L for a subjective trustee corresponds to the weighted sum of the trustee's utility when the state is $\omega = L$ and the trustee knows it with probability (with probability ps), when the state is $\omega = L$ and the trustee knows it with probability (with probability $(1-p)s$), and when the trustee is uncertain about the state (with probability $1 - s$).

$$Eu_{LH} = ps \cdot \hat{u}_i(\Phi_L) + (1-p)s \cdot \hat{u}_i(\Phi_H) + (1-s) \cdot \hat{u}_i(\Phi_U) \tag{3.24}$$

First, we focus on the case of guilt-averse trustees. To conclude from the equations below, recall that (i) since $\Phi_L < \Phi_U < \Phi_H$, it follows that $\hat{u}_{g,i}(\Phi_L) > \hat{u}_{g,i}(\Phi_U) > \hat{u}_{g,i}(\Phi_H)$, and (ii) p and $s \in [0,1]$.

$$\begin{aligned} Eu_L - Eu_H &= ps \cdot \hat{u}_{g,i}(\Phi_L) + (1-ps) \cdot \hat{u}_{g,i}(\Phi_U) - (1-p)s \cdot \hat{u}_{g,i}(\Phi_H) - (1-s+ps) \cdot \hat{u}_{g,i}(\Phi_U) \\ &= ps \cdot \hat{u}_{g,i}(\Phi_L) + (1-ps-1+s-ps) \cdot \hat{u}_{g,i}(\Phi_U) - (1-p)s \cdot \hat{u}_{g,i}(\Phi_H) \\ &= ps \cdot \underbrace{[\hat{u}_{g,i}(\Phi_L) - \hat{u}_{g,i}(\Phi_U)]}_{>0} + (1-p)s \cdot \underbrace{[\hat{u}_{g,i}(\Phi_U) - \hat{u}_{g,i}(\Phi_H)]}_{>0} \end{aligned} \tag{3.25}$$

$$\begin{aligned} Eu_L - Eu_{LH} &= ps \cdot \hat{u}_{g,i}(\Phi_L) + (1-ps) \cdot \hat{u}_{g,i}(\Phi_U) - ps \cdot \hat{u}_{g,i}(\Phi_L) - (1-p)s \cdot \hat{u}_{g,i}(\Phi_H) - (1-s) \cdot \hat{u}_{g,i}(\Phi_U) \\ &= (ps-ps) \cdot \hat{u}_{g,i}(\Phi_L) + (1-ps-1+s) \cdot \hat{u}_{g,i}(\Phi_U) - (1-p)s \cdot \hat{u}_{g,i}(\Phi_H) \\ &= s(1-p) \cdot \underbrace{[\hat{u}_{g,i}(\Phi_U) - \hat{u}_{g,i}(\Phi_H)]}_{>0} > 0 \end{aligned} \tag{3.26}$$

We can conclude that, under uncertainty, a subjective guilt-averse trustee who follows a coarse mapping, will acquire signal S_L, but neither signal S_H not both signals.

Second, we focus on the case of reciprocal trustees. To conclude from the equations below, recall that (i) since $\Phi_L < \Phi_U < \Phi_H$, it follows that $\hat{u}_{r,i}(\Phi_L) < \hat{u}_{r,i}(\Phi_U) < \hat{u}_{r,i}(\Phi_H)$, and (ii) p and $s \in [0,1]$.

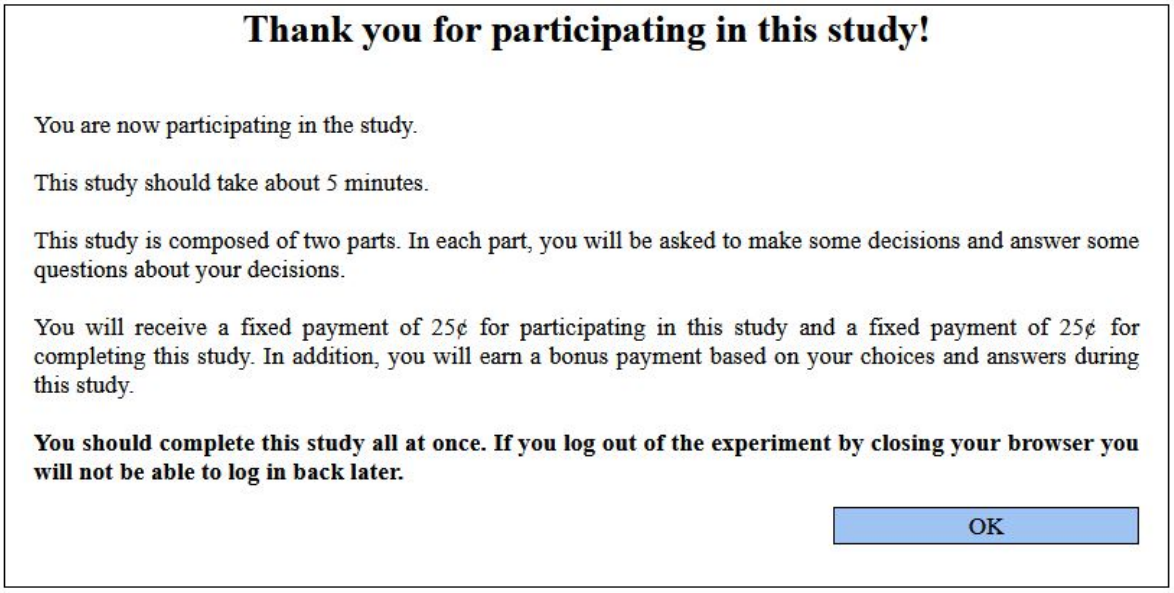
$$\begin{aligned} Eu_H - Eu_L &= (1-p)s \cdot \hat{u}_{r,i}(\Phi_H) + (1-s+ps) \cdot \hat{u}_{r,i}(\Phi_U) - ps \cdot \hat{u}_{r,i}(\Phi_L) - (1-ps) \cdot \hat{u}_{r,i}(\Phi_U) \\ &= (s-ps) \cdot \hat{u}_{r,i}(\Phi_H) + (1-s+ps-1+ps) \cdot \hat{u}_{r,i}(\Phi_U) - ps \cdot \hat{u}_{r,i}(\Phi_L) \\ &= (s-ps) \cdot \underbrace{[\hat{u}_{r,i}(\Phi_H) - \hat{u}_{r,i}(\Phi_U)]}_{>0} + ps \cdot \underbrace{[\hat{u}_{r,i}(\Phi_U) - \hat{u}_{r,i}(\Phi_L)]}_{>0} \end{aligned} \tag{3.27}$$

$$\begin{aligned} Eu_H - Eu_{LH} &= (1-p)s \cdot \hat{u}_{r,i}(\Phi_H) + (1-s+ps) \cdot \hat{u}_{r,i}(\Phi_U) - ps \cdot \hat{u}_{r,i}(\Phi_L) \\ &\quad -(1-p)s \cdot \hat{u}_{r,i}(\Phi_H) - (1-s) \cdot \hat{u}_{r,i}(\Phi_U) \\ &= (1-s+ps-1+s) \cdot \hat{u}_{r,i}(\Phi_U) - ps \cdot \hat{u}_{r,i}(\Phi_L) \\ &= ps \cdot \underbrace{[\hat{u}_{r,i}(\Phi_U) - \hat{u}_{r,i}(\Phi_L)]}_{>0} \end{aligned} \tag{3.28}$$

We can conclude that, under uncertainty, a subjective reciprocal trustee who follows a coarse mapping, will acquire signal S_H, but neither signal S_L not both signals.

3.B Screens from the online experiment

3.B.1 Trustors' screens



Part 1: Instructions

All payments in this part are bonus payments, they do not include your fixed payment.

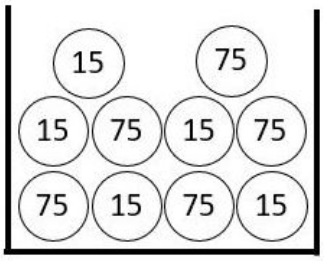
In this part of the study, you can have one of two following roles: Participant A or Participant B.
You have been randomly selected to be **Participant A**.
You will be matched with a fellow Mturk worker who has been randomly selected to be Participant B.

Your role is to choose between two options (orange or green) that have different consequences on both your earnings and the earnings of Participant B.

Before you make a decision, the computer program randomly draws a ball from an urn containing ten balls. Out of these ten balls, five are marked with the number 15 and five are marked with the number 75.

The number marked on the ball determines your earnings in the orange option.

- If the number on the ball is 15, you earn 15¢.
- If the number on the ball is 75, you earn 75¢.



Participant B knows the consequences of the number marked on the ball on your earnings.

More details on the consequences of the two options (orange or green) are given on the next screen.

OK

Part 1: Instructions

The consequences of your choice of color are explained below.





If you choose the orange button, the amount you earn depends on the number on the ball. If the ball is marked with 15, you earn 15¢. If the ball is marked with 75, you earn 75¢. Participant B earns 90¢, regardless of the number marked on the ball.

If you choose the green button, you entrust Participant B with 200¢ to allocate between the two of you, regardless of the number marked on the ball.

Participant B knows the orange button yields fixed earnings while the green button leaves him/her to decide on how to allocate 200¢.

OK

Part 1: Comprehension Questionnaire

Click to remind me of the instructions

To make sure that the instructions are clear, please answer the following questions. If you have a doubt about the instructions, click on the button at the top of your screen.

Suppose that you choose the green button, and that Participant B chooses to send you 60¢.

1) How much do you earn from this decision?
[ ] ¢

2) How much does Participant B earn from this decision?
[ ] ¢

Suppose that you choose the orange button.

3) How much do you earn from this decision if the number on the ball is 15?
[ ] ¢

4) How much do you earn from this decision if the number on the ball is 75?
[ ] ¢

5) How much does Participant B earn from this decision?
[ ] ¢

Click to check my answers

Part 1: Guess Participant B's decision

Click to remind me of the instructions

You now have the opportunity to earn an additional 50¢ if you make a correct guess.

If you choose the green button, Participant B decides how to allocate 200¢ between the two of you (by increments of 15¢). We would like to know **how much you expect Participant B to send you** in this case.

Participant B decides how to allocate 200¢ between the two of you, assuming that you choose the green button. Your answer will be compared with Participant B's decision. If you guess the amount correctly (plus or minus 15¢), you will earn 50¢ in addition to your other earnings.

If the number on the ball is 15, **you have to give up 15¢** to let Participant B decide the final earnings (green button). How much do you expect Participant B to send you in this case?

-- select an option -- ¢

If the number on the ball is 75, **you have to give up 75¢** to let Participant B decide the final earnings (green button). How much do you expect Participant B to send you in this case?

-- select an option -- ¢

OK

Part 1: Your decision

The computer program randomly selected **a ball marked with the number 15.**

You reported that **you expect that Participant B will send you 30¢** if you choose green .

Click on one of the buttons below to make a decision:





If you choose the orange button, you earn 15¢ and Participant B earns 90¢.

If you choose the green button, you entrust Participant B with 200¢ to allocate between the two of you.

OK

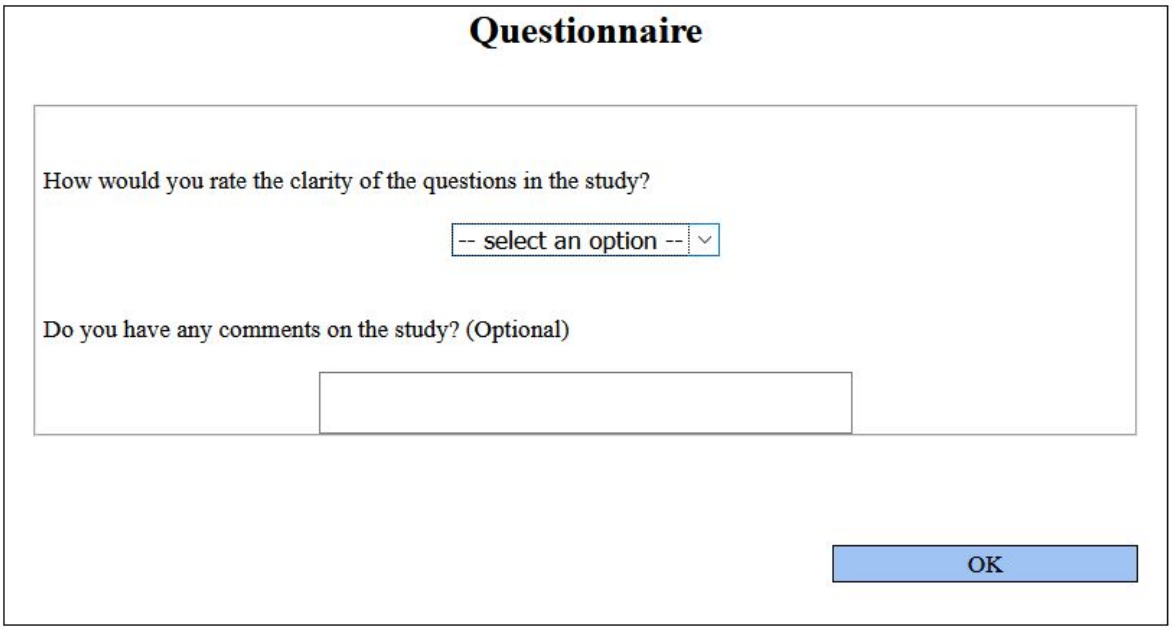
Part 2

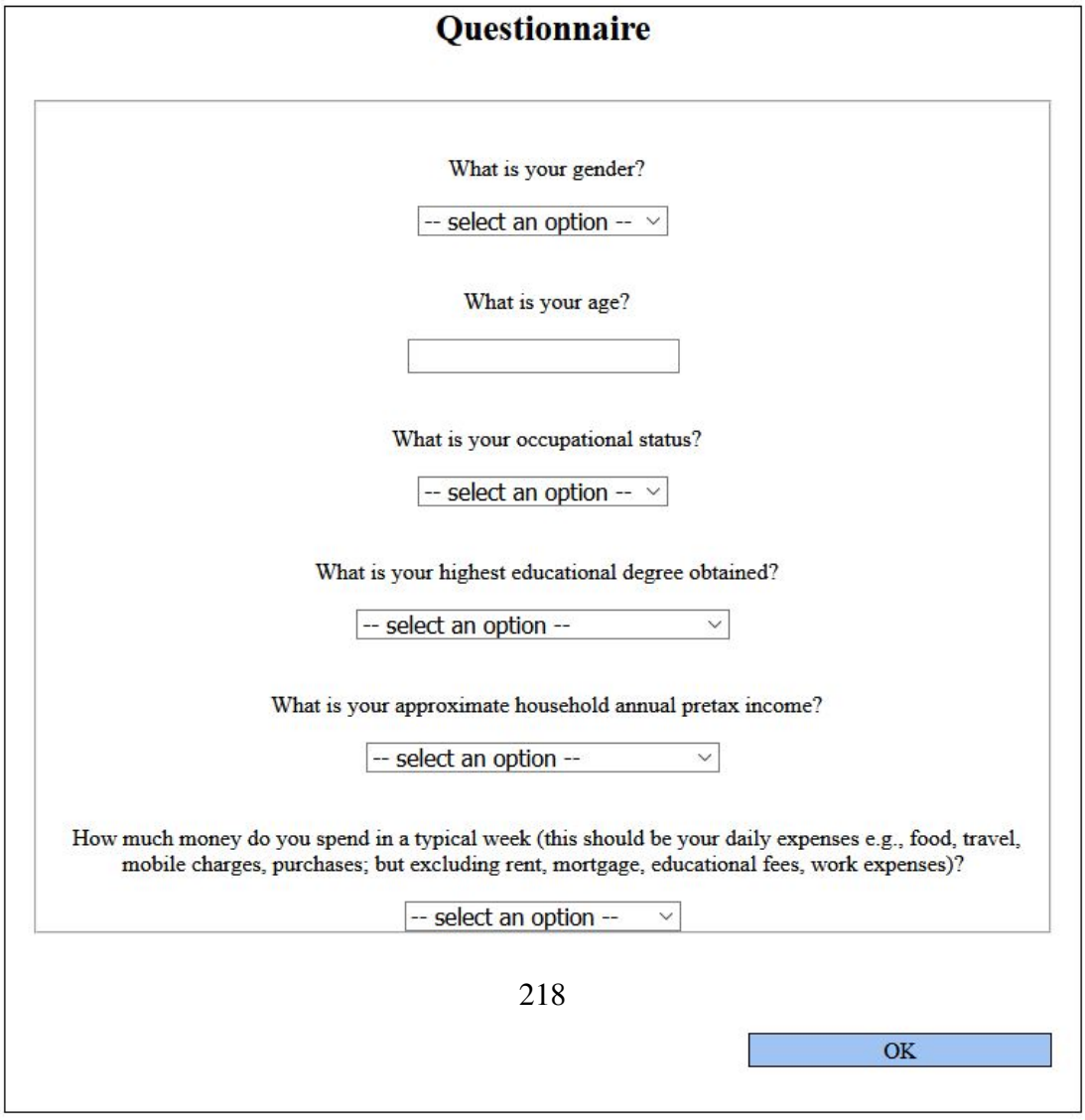
All payments in this part are bonus payments, they do not include your fixed payment.

In this part of the study, you and all the other participants to the study must guess a number between 0 and 100 (inclusive). The participant whose chosen number is the closest to the two-third of the mean of all chosen numbers earns 100¢. In case of a tie, the participant who earns the 100¢ will be chosen at random by the computer program.

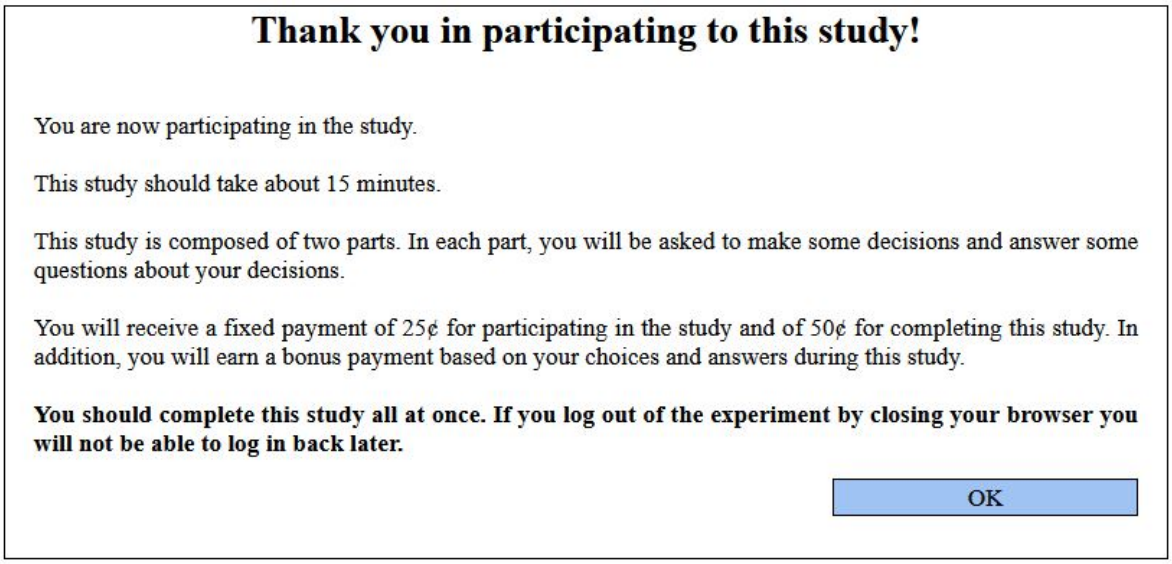
What is your chosen number?

OK


Questionnaire
How would you rate the clarity of the questions in the study?
-- select an option --
Do you have any comments on the study? (Optional)
OK


Questionnaire
What is your gender?
-- select an option --
What is your age?
What is your occupational status?
-- select an option --
What is your highest educational degree obtained?
-- select an option --
What is your approximate household annual pretax income?
-- select an option --
How much money do you spend in a typical week (this should be your daily expenses e.g., food, travel, mobile charges, purchases; but excluding rent, mortgage, educational fees, work expenses)?
-- select an option --
OK

3.B.2 Trustees' screens



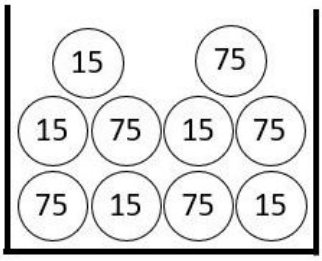
Part 1: Instructions

All payments in this part are bonus payments, they do not include your fixed payment.

In this part of the study, you can have one of two following roles: Participant A or Participant B.
You have been randomly selected to be **Participant B**.
You will be matched with a fellow Mturk worker who has been randomly selected to be Participant A.

The role of Participant A is to choose between two options (orange or green) that have different consequences on both your earnings and the earnings of Participant A.

Before Participant A makes a decision, the computer program randomly draws a ball from an urn containing ten balls. Out of these ten balls, five are marked with the number 15 and five are marked with the number 75. The number marked on the ball determines the earnings of Participant A in the orange option.



More details on the consequences of the two options (orange or green) are given on the next screen.

OK

Part 1: Instructions

Participant A's options are presented below.





If Participant A chooses orange, **you have no choice to make**. The earnings of Participant A are determined by the number marked on the ball. If the ball is marked with 15, Participant A earns 15¢. If the ball is marked with 75, Participant A earns 75¢. You earn 90¢, regardless of the number marked on the ball.

If Participant A chooses green, **you have to make a choice that impacts both of your earnings**. Participant A entrusts you with 200¢ and you choose how to allocate these 200¢ between the two of you, regardless of the number marked on the ball.

Participant A receives the same instructions and, in addition, he/she knows which ball has been drawn before making a decision. In contrast, you will not know which ball has been drawn before making your decision.

Participant A's decision is not influenced by your choices. You will not be informed of Participant A's decision when making your own decision.

OK

Part 1: Comprehension Questionnaire

Click to remind me of the instructions

To make sure that the instructions are clear, please answer the following questions. If you have a doubt about the instructions, click on the button at the top of your screen.

Suppose that Participant A chooses the green button, and you choose to send 60¢ to Participant A.

1) How much do you earn from this decision?
[ ] ¢

2) How much does Participant A earn from this decision?
[ ] ¢

Suppose that Participant A chooses the orange button.

3) How much do you earn from this decision?
[ ] ¢

4) If the ball drawn by the computer program is marked with the number 15, how much does Participant A earn from this decision?
[ ] ¢

5) If the ball drawn by the computer program is marked with the number 75, how much does Participant A earn from this decision?
[ ] ¢

Click to check my answers

Part 1: Guess Participant A's expectations

Click to remind me of the instructions

You now have the opportunity to earn an additional 50¢ if you make the correct guess.

We ask Participant A to guess how much he/she expects to receive from you, if he/she chooses the green button. We would like to know how much you think Participant A expects to receive from you.

Your answer will be compared with the actual expectation of participant A. If you have guessed the amount correctly (plus or minus 15¢), you will earn 50¢ in addition to your other earnings.

Please pay attention to the different scenarios to answer the following questions.

Your guesses if Participant A chooses the green button:

If the number on the ball is 15, **Participant A has to give up 15¢** to let you decide the final earnings (green button). How much do you think Participant A expects to receive from you (by increments of 15¢) in this case?

-- select an option -- ¢

If the number on the ball is 75, **Participant A has to give up 75¢** to let you decide the final earnings (green button). How much do you think Participant A expects to receive from you (by increments of 15¢) in this case?

-- select an option -- ¢

At the end of the experiment, the answer corresponding to the true number marked on the ball will be selected for payment.

OK

Part 1: Your decisions

Click to remind me of the instructions

If Participant A chooses the green button, he/she entrusts you with 200¢. In this case, you have to decide how much ¢ to send to Participant A out of these 200¢ in two cases:

- **If you learn that the number on the ball is 15** (Decision 15).
- **If you learn that the number on the ball is 75** (Decision 75).

If you remain uninformed about the number marked on the ball, Participant A will receive the average of the amount indicated in Decision 15 and the amount indicated in Decision 75.

Decision 15:

If the number on the ball is 15, you reported that **Participant A expects to receive 45¢**. How much would you send to Participant A in this case?

-- select an option -- ¢

Decision 75:

If the number on the ball is 75, you reported that **Participant A expects to receive 90¢**. How much would you send to Participant A in this case?

-- select an option -- ¢

OK

Part 1: Instructions

You have now the possibility to be informed about the number marked on the ball.

- If you learn that the number on the ball is 15, Decision 15 is implemented: you keep 140¢.
- If you learn that the number on the ball is 75, Decision 75 is implemented: you keep 95¢.

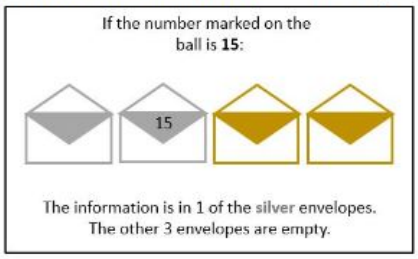
You can also remain uninformed about the number marked on the ball:

- Then, the average of Decision 15 and Decision 75 is implemented: you keep 117.5¢.

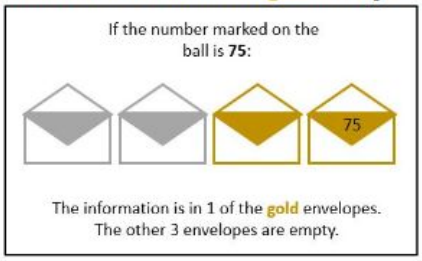
The information is hidden in one of 4 envelopes. There are 2 silver envelopes and 2 gold envelopes.

One of these envelopes contains the number marked on the ball (either 15 or 75) and the other three envelopes are empty.

If the ball is marked with the number 15, the information is in one of the two silver envelopes.



If the ball is marked with the number 75, the information is in one of the two gold envelopes.



You will have to choose between three possibilities. You can choose to open:

- 1 silver envelope
- 1 gold envelope
- 1 silver envelope and 1 gold envelope

Your choice does not affect Participant A's decision.

OK

Part 1: Instructions

Click to remind me of the instructions

If you choose to open only 1 silver envelope, you will either:

- learn that the number on the ball is a 15: you keep 140¢ (Decision 15)
- remain uninformed about the number on the ball: you keep 117.5¢ (average of Decisions 15 and 75)

If you choose to open only 1 gold envelope, you will either:

- learn that the number on the ball is a 75: you keep 95¢ (Decision 75)
- remain uninformed about the number on the ball: you keep 117.5¢ (average of Decisions 15 and 75)

If you choose to open 1 silver envelope and 1 gold envelope, you will either:

- learn that the number on the ball is a 15: you keep 140¢ (Decision 15)
- learn that the number on the ball is a 75: you keep 95¢ (Decision 75)
- remain uninformed about the number on the ball: you keep 117.5¢ (average of Decisions 15 and 75)

OK

Part 1: Comprehension Questionnaire

Click to remind me of the instructions

To make sure that the instructions are clear, please answer the following questions. If you have a doubt about the instructions, click on the button at the top of your screen.

1) What happens if you learn that the number on the ball is a 15?

- ○ Decision 15 is implemented.
- ○ Decision 75 is implemented.
- ○ The average of Decision 15 and Decision 75 is implemented.

2) What happens if you remain uninformed about the number marked on the ball?

- ○ Decision 15 is implemented.
- ○ Decision 75 is implemented.
- ○ The average of Decision 15 and Decision 75 is implemented.

3) What happens if you open a silver envelope?

- ○ You either learn that the number on the ball is 15 or remain uninformed.
- ○ You either learn that the number on the ball is 75 or remain uninformed.

4) What happens if you open both a silver and a gold envelopes?

- ○ You learn for sure whether the number on the ball is 15 or 75.
- ○ You either learn that the number on the ball is 15, 75 or remain uninformed.

Click to check my answers

Part 1: Your decision

Click to remind me of the instructions

Below are 2 silver envelopes and 2 gold envelopes.
You can now choose to open:

- 1 silver envelope
- 1 gold envelope
- 1 silver envelope and 1 gold envelope

Please pay attention to your choice since you will be asked to explain it at the end of the study.

Click on the envelope(s) that you wish to open. Click again on an envelope if you want to unselect it.
You can try out several choices before your final choice.



You chose to open 1 gold envelope. You will either:

- learn that the number on the ball is a 75 and keep 95¢ (Decision 75).
- remain uninformed and keep 117.5¢ (average of Decisions 15 and 75).

OK

Part 1: Feedback



The envelope(s) you opened were empty. You remain uninformed of the ball drawn by the computer program.

Therefore, if Participant A chooses the green button, Participant A will receive the average between the amount you chose to send him or her in Decision 15 and the amount you chose to send him or her in Decision 75, that is (60+105)/2 = 82.5¢ and you will keep 117.5¢.

If Participant A chooses the orange button, you will receive 90¢. Participant A will receive 15¢ if the number on the ball is 15 and 75¢ if the number on the ball is 75.

OK

Part 2

All payments in this part are bonus payments, they do not include your fixed payment.

In this part of the study, you and all the other participants to the study must guess a number between 0 and 100 (inclusive). The participant whose chosen number is the closest to the two-third of the mean of all chosen numbers earns 100¢. In case of a tie, the participant who earns the 100¢ will be chosen at random by the computer program.

What is your chosen number?

OK

Questionnaire

In Part 1 of the study, how likely is it that Participant A chooses the green button (letting you decide the final earnings) if the number on ball was 15?

-- select an option --

In Part 1 of the study, how likely is it that Participant A chooses the green button (letting you decide the final earnings) if the number on ball was 75?

-- select an option --

In Part 1 of the study you were given the opportunity to receive information about the ball drawn by the computer program. The information was contained in one of four envelopes (2 silver and 2 gold). You chose to open 1 envelope(s). Please briefly explain why. You will earn 50¢ to answer this question. (Optional)

How would you rate the clarity of the questions in the study?

-- select an option --

Do you have any comments on the study? (Optional)

OK

Questionnaire

What is your gender?

-- select an option --

What is your age?

What is your occupational status?

-- select an option --

What is your highest educational degree obtained?

-- select an option --

What is your approximate household annual pretax income?

-- select an option --

How much money do you spend in a typical week (this should be your daily expenses e.g., food, travel, mobile charges, purchases; but excluding rent, mortgage, educational fees, work expenses)?

-- select an option --

OK

Thank you for participating in this study!

You have now completed the study.

Thank you!

You will receive your fixed payment within 48 hours and your bonus payment (i.e, the sum of earnings from the two parts) within a week from today.

If you have any questions concerning this study, you can contact us at rimbaud@gate.cnrs.fr

Your confirmation code to be entered on Mturk webpage is your Mturk worker ID. Please submit the HIT on Mturk with this ID.

You can close this window now.

3.C Additional Results

3.C.1 Trustor's behavior

We showed in Section 3.5.1 that trustor's expect to receive more from the trustees when their outside option is high rather than low. Consistent with Equation 3.1, 85.45% of trustors who choose to go *In* expects to receive more than their outside option. Moreover, the share of trustees choosing *In* is lower when the outside option is High (51.85%) rather than Low (86.08%) (chi-square test, $p < 0.001$).

3.C.2 Trustees' justification of their sampling strategies

We classified the participants' justification of their sampling strategies in four categories (excluding 11 trustees who did not fill in this optional question). The first category pools the trustees who made their choice out of curiosity, *e.g.*, "*I was just curious to see if I would find a 15 or 75*". Second, we grouped together participants who mentioned their intention to maximize their payoff, *e.g.*, "*I chose to open 1 silver envelope hoping it would contain a 15 and then I would maximize my earnings*". In the third category, we pooled the participants who reported having made their choice at random, *e.g.*, "*I chose 1 envelope honestly just based on feeling*". The last category contains answers that we could not classify in the other three categories.

Table 3.C.1a shows that when opening one envelope only, the majority of belief-dependent trustees choose at random; while they are motivated by curiosity when they open both envelopes. Table 3.C.1b shows that the majority opened a silver envelope to maximize their payoff, while they opened both envelope to satisfy their curiosity.

Table 3.C.1: Trustees' justification of their sampling strategies

(a) Belief-independent trustees

	Curiosity	Payoff	Random	Other	Total (n)
Open Silver	12.50%	25.00%	50.00%	12.50%	8
Open Gold	36.36%	9.09%	36.36%	18.18%	11
Open Both	62.50%	4.17%	20.83%	12.50%	24

(b) Guilt averse trustees

	Curiosity	Payoff	Random	Other	Total (n)
Open Silver	4.17%	79.17%	4.17%	12.50%	24
Open Gold	12.50%	37.50%	12.50%	37.50%	8
Open Both	75.00%	12.50%	0.00%	12.50%	8

(c) Reciprocal trustees

	Curiosity	Payoff	Random	Other	Total (n)
Open Gold	0.00%	25.00%	0.00%	75.00%	4

3.C.3 Trustees' likelihood of having a given sampling strategy

Table 3.C.2 reports the average marginal effect of a multinomial logit model using a categorical variable equals to 0 if the trustee opened a silver envelope only, 1 if the trustee opened a gold envelop only and 2 if the trustee opened both a silver and a gold envelopes as the dependent variable. Regressors include a dummy variable equal to 1 if the trustee is guilt-averse, and 0 if the trustee is belief-independent. Guilt-averse trustees are more likely to open a silver envelope and less likely to open both envelopes than belief-independent trustees and the results are significant at the 0.1% level. These results are robust to the inclusion of individual controls.

Table 3.C.2: Average marginal effects of preferences types on the likelihood of each sampling strategy

	Open Silver		Open Gold		Open both	
Belief-independent	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
Guilt averse	0.428***	0.453***	-0.108	-0.105	-0.320***	-0.348***
	(0.092)	(0.086)	(0.087)	(0.085)	(0.093)	(0.085)
Ind. controls	No	Yes	No	Yes	No	Yes
Observations	94	94	94	94	94	94

Notes: This Table reports the average marginal effects estimated by Multinomial Logit models. Individual controls include the amount guessed in the beauty contest game, and socio-demographic characteristics (age, gender, annual income, weekly expenditure). Standard errors are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

3.C.4 Determinants of beliefs, returns and preference type.

To investigate the determinants of participants' beliefs, we estimated a linear regression of the difference in beliefs for both trustors and trustees on participants' individual characteristics. The OLS coefficients are displayed in columns (1) and (2) in Table Table 3.C.3, respectively. We find that an increase in the perceived clarity of the instructions increases trustor's sensitivity to our treatment manipulation ($p = 0.038$), but not trustees'. Surprisingly, we find no effect of participants' guess in the beauty contest on their sensitivity to the treatment manipulation ($p = 0.557$ and $p = 0.536$).

In addition, we investigate the determinants of trustees' difference in conditional return choices. To do so, we estimated a linear regression of the difference in conditional returns on trustees' individual characteristics. The OLS coefficients are displayed in column (3). We find no effect of trustees' individual characteristics on their conditional return choices.

Finally, we investigate the determinants of trustees' preference type. To do so, we estimated logit regression of the likelihood of having belief-independent or guilt-averse preferences on trustees' individual characteristics. The average marginal effects are displayed in columns (4) and (5), respectively. We find no effect of trustees' individual characteristics on their preference type.

Table 3.C.3: Determinants of participants' beliefs, trustees' conditional return decisions and preferences type.

Dep. var:	Diff. belief trustors	Diff. belief trustees	Diff. return trustees	Types Trustees	
				Belief Ind.	Guilt-averse
	(1)	(2)	(3)	(4)	(5)
Level of reasoning	0.078	-0.077	0.141	0.001	-0.002
	(0.133)	(0.125)	(0.135)	(0.003)	(0.003)
Female	-2.502	5.807	1.464	-0.051	0.019
	(4.933)	(5.198)	(5.169)	(0.112)	(0.111)
Age	-0.226	-0.321	-0.106	0.008	-0.005
	(0.209)	(0.212)	(0.238)	(0.005)	(0.005)
Annual income	0.306	-1.551	0.226	0.0108	-0.003
	(1.975)	(1.957)	(1.986)	(0.043)	(0.042)
Weekly expenditure	-4.128	6.009	-0.315	0.002	-0.016
	(3.629)	(3.277)	(3.263)	(0.070)	(0.069)
Clarity instructions	8.619*	2.721	-2.137	0.122	-0.110
	(4.114)	(3.621)	(3.771)	(0.083)	(0.081)
Constant	42.19***	31.67*	21.79	–	–
	(13.38)	(14.17)	(14.34)		
Observations	160	160	98	98	98

Notes: Table Table 3.C.3 displays the OLS coefficients of participants' individual characteristics on trustors' (column (1)) and trustees' (column (2)) differences in beliefs between the Low and the High game, trustees' differences in return between the Low and the High game (column (3)), as well as the marginal effect from a logit regression of trustees' individual characteristics on their preference type (columns (5) and (6)). Standard errors are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

3.D Robustness checks Restricted sample

We pre-registered that we will verify the robustness of our findings by excluding from the analyses participants who indicated in the post-experimental questionnaire that (i) the instructions were not extremely clear or that (ii) the had trouble understanding the instructions.

59 trustors and 81 trustees indicated that the instructions were not extremely clear (43.75% of participants). In addition, 13 trustees indicated that they encountered comprehension problem with the instructions while indicating that the instructions were extremely clear (4.06% of participants). In the following section, we excluded these participants from the analyses.

3.D.1 Are beliefs affected by the outside option manipulation?

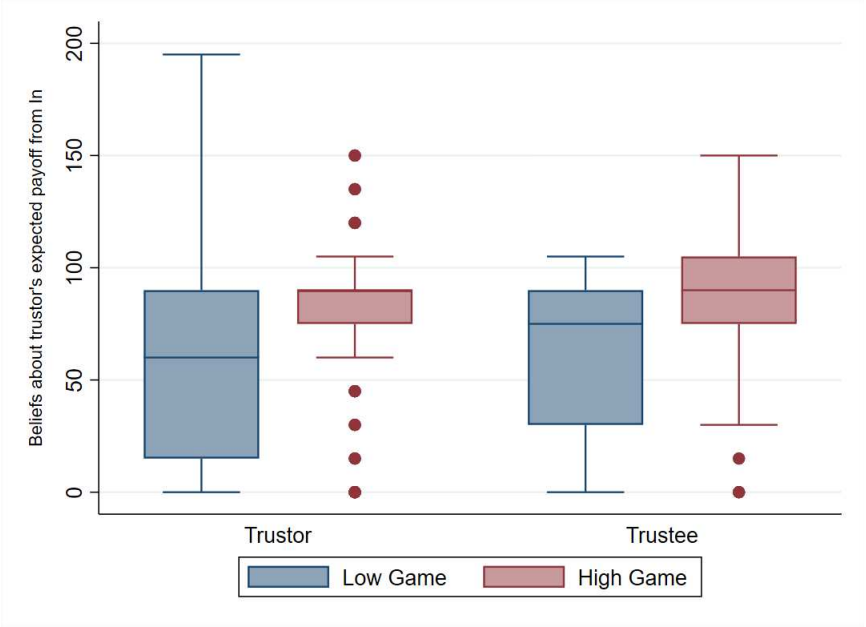
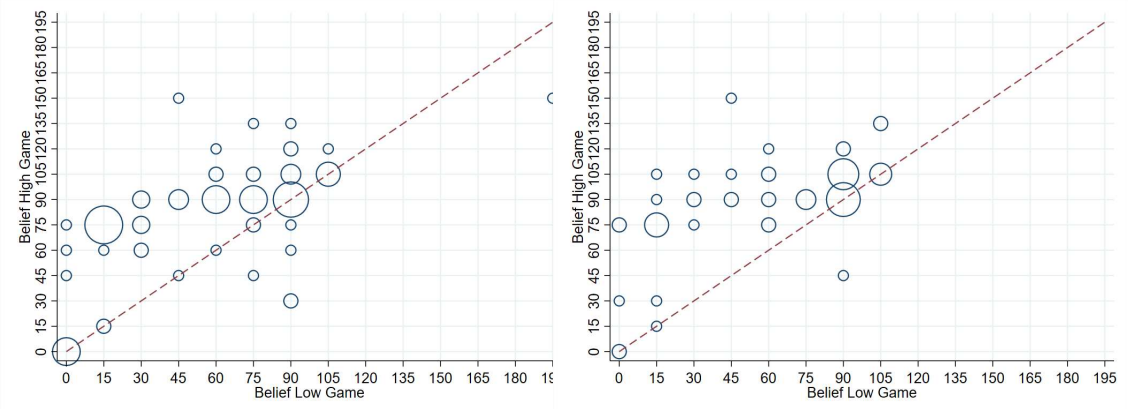


Figure 3.D.1: Distribution of trustors and trustees' beliefs about trustors' payoff from *In*.

The median and interquartile range of the distribution of beliefs remains the same as in the main text in both games. Result 1 is not affected by participants comprehension of the experimental instructions.



(a) Trustors' first-order beliefs **(b)** Trustees' second-order beliefs

Figure 3.D.2: Distribution of individual beliefs about trustors' expected payoff from *In*

Turning to the individual level, the proportion of participants who verifies Auxiliary Hypothesis 1 increases slightly. 62% (vs. 53.13%) of trustors and 68.18% (vs. 61.25%) of trustees hold higher beliefs in the High game than in the Low game.

In contrast, 5.94% (vs. 10%) of trustors and 1.52% (vs. 8.13%) of trustees indicated higher beliefs in the Low game than in the high Game, while 32.67% (vs. 28.75%) of trustors and 30.30% (vs. 38.75%) of trustees indicated similar expectations regardless of the game being played.

3.D.2 Are trustees motivated by belief-dependent preferences?

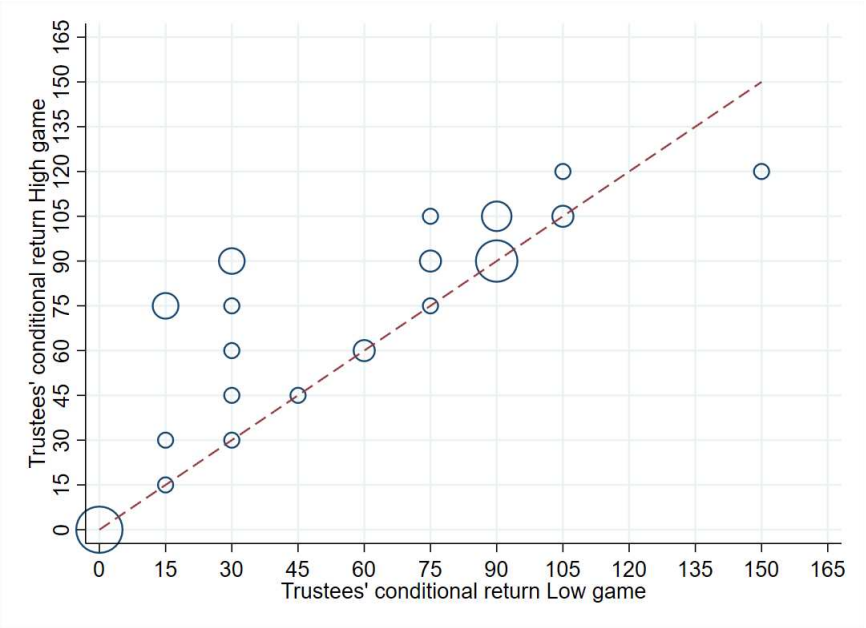


Figure 3.D.3: Trustees' return strategies

40% (n=18; vs. 43.88%, n = 43) of trustees can be classified as guilt-averse. Only one trustee can be classified as reciprocal (vs. 4.08%, n = 4). Result 2 is not affected by participants comprehension of the experimental instructions.

3.D.3 How do belief-based preferences affect information acquisition

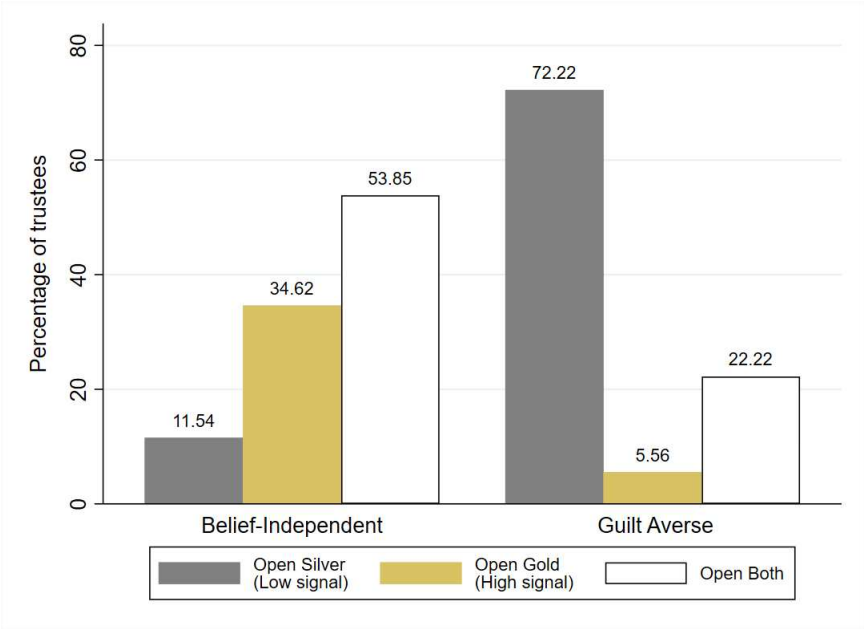


Figure 3.D.4: Information acquisition strategy.

Conclusion

The aim of this thesis was to question the scope of guilt-aversion influence on behavior. We showed that its range of influence is broader than initially suspected but it is a less robust phenomenon than the previous literature suggested. On the one hand, we proposed a model and experimentally demonstrated that the range of influence of guilt extends toward players who are not affected monetarily by the decision-maker's action. On the other hand, we revealed that individuals develop strategies for acquiring information about others' expectations that allow them to implement the least pro-social option without suffering the psychological cost of guilt aversion.

In Chapter 1, we considered a situation where a donor can send money to a recipient. As in the case of charitable donations or government transfers to developing countries, this money must be transmitted through an intermediary who can embezzle some of it. We tested whether the direction of the guilt (toward the recipient or toward the donor) affects the intensity or the prevalence of guilt-aversion among intermediaries. Our experimental results indicated that the proportion of intermediaries who experience guilt was similar regardless of whether the guilt is toward the recipient or toward the donor. The absence of difference across treatments was confirmed when looking at the intensity of their aversions. This is striking because embezzlement affects the earnings of the recipient but not those of the donor. It shows that the mechanisms at play in

guilt aversion extend in situations where decisions have no direct monetary consequences.

In this chapter, we measured guilt aversion toward the donor and toward the recipient in two separate treatments. Given that we are the first to document the existence of guilt toward the donor, the literature is agnostic about the joint effect of these two types of expectations. Therefore, an important extension of this study would be to test a treatment in which intermediaries would be informed about both donors' and recipients' expectations: whether one effect overrides the other, or whether their effects are cumulative. This would lead to a complex design, though. Furthermore, extending the present study by including several rounds of play, instead of a one-shot game, could allow us to test Balafoutas (2011) hypothesis of a vicious circle of corrupt norms. If donors or recipients expect a high level of embezzlement in a group, intermediaries can embezzle without feeling guilty, which in turn increases the expectations of embezzlement.

In Chapter 2, we assessed whether guilt aversion is moderated by the vulnerability of the player towards whom the decision-maker may feel guilty. To do so, we designed four novel Quasi-Trust mini-games where we varied systematically the vulnerability of the co-players. We found that neither the proportion of guilt-averse second movers nor the intensity of their guilt aversion differed significantly across the four games (*i.e.* the four combinations of vulnerability), and across the two treatments (*i.e.* guilt elicited toward an active *vs.* a passive player). In particular, second movers exhibited a guilt-averse behavior even toward the beliefs of players who were not vulnerable at all. Interestingly, this confirms the relevance of guilt aversion in situations of investment and it reveals its importance in situations of donation or exploitation.

To sum up, we have shown that the vulnerability of co-players does not affect guilt aversion. The insensitivity of players B's guilt aversion to manipulations of the co-

player's payoffs and intentions could, however, be interpreted as a sign of confusion by our subjects (i.e., subjects did not understood the different games). Yet, players A's behavior pleads against this interpretation. We found their behavior to be game-dependent, in line with our model of lexicographic-altruism. Alternatively, the display of players B's choices may have reduced the potential impact of the co-players vulnerability. Indeed, participants were asked to make their choice conditional on four levels of expectations of the other player. This contextualization of choices, traditional when testing belief-based preferences, has potentially overcome the information provided when introducing the game that was supposed to trigger a reaction based on the other player's vulnerability. This alternative account of our results could be tested by informing of the co-player expectations at the beginning and asking players B to condition their choices on the different manipulations of their co-players' vulnerability.

In Chapter 3, we investigated whether individuals with belief-dependent preferences self-servingly bias their information acquisition strategy in order to minimise the tension between their monetary interest and their belief-dependent motivation. We tested our predictions in on online experiment where we gave the opportunity to trustees to acquire information about the trustors' outside option, and thereby about the trustors' expectations. Our results highlight that this impact of guilt aversion depends on the (un)certainty about others' expectations. Our main contribution is to show that individuals can bias their information acquisition strategy toward self-serving signals to avoid paying the expected monetary cost of following their conscience, i.e. making the choice that corresponds to the true state of the world. In the mainstream literature on belief-dependent preferences, beliefs are perfectly observable at the time where the actions are implemented. Our findings suggest that previous results on the positive impact of belief-dependent preferences on pro-social choices capture an upper bound of this impact.

It is noteworthy that both our theoretical model and our experimental design consider the case of a coarse mapping of beliefs, that is, the belief about the state of the world is a step function. If we were to relax this feature and allow for linear beliefs, the optimal choice of a trustee with “illusory" belief-dependent preferences would be to avoid all information, a choice that is not possible in our experiment. Therefore, a natural extension would be to consider a linear mapping of beliefs in an experimental set-up which would allow for pure information avoidance. However, this extension requires that participants to be able to update their beliefs in a Bayesian manner, which has been shown to be quite difficult (e.g., Grether, 1980; Belot et al., 2012). Furthermore, one can wonder if information acquisition strategies are influenced by the “default option”. In the present study, participants had to select the information they wanted to acquire. What would happen if, by default, all information were selected, and participants had had to un-select the information they did not want to acquire? Grossman and Van Der Weele (2017) results suggest that this alternative design may lead to less strategic information acquisition.

In terms of public policy, the results from this thesis suggest to modify the communication in anti-corruption campaigns of information. Public communication (Reinikka and Svensson, 2004) usually focus on the potential recipients’ expectations. However, anti-corruption policies should publicize the high expectations not only of recipients but also of donors to limit embezzlement by intermediaries. More broadly, although our data reinforce the relevance of public campaigns aiming to nudge behaviors through the mechanisms of guilt aversion, they also suggest that it is not enough to “simply” increase the availability of information about others’ expectations but it is also needed to ensure that this information cannot be overlooked. Indeed, in case of uncertainty regarding the beliefs of others, a majority of belief-dependent individuals seek self-serving information, which eventually leads them to select the payoff-maximising action without

compromising their belief-dependent motivation.

To date, economists have focused on the anticipation of guilt, which, if aversive, motivates prosocial behaviors. Yet, psychologists have highlighted that the experience of guilt can also encourage prosocial behavior (Baumeister et al., 1994). Indeed, the experience of guilt can motivate reparative behavior (e.g., Ketelaar and Tung Au, 2003). Considering this aspect of guilt could enrich the economic literature on excuses that has focused on apologies triggered by the harm done, rather than by the disappointment of expectations (e.g., Abeler et al., 2010). Furthermore, the display of guilt can appease victims or bystanders.

Experiments in Chapters 1 and 2 were conducted in the lab while, in Chapter 3, we implemented an online experiment on Amazon Mechanical Turk due to the world pandemic of Covid-19. This situation allowed us to test the existence of guilt-averse behaviors in a context that increases social distance. In contrast with the work of Morell (2019) suggesting that social distance limits guilt-aversion, we do observe guilt-averse individuals in our sample. It suggests that guilt aversion exists in a wider range of situations than those usually investigated (see also Bellemare et al., 2011 for the only other online study on guilt aversion). This opens the way to new paths of research to study guilt with much larger samples.

Both for lab and online experiments, one can wonder about the external validity of their results. We are aware of only one paper which studies how guilt aversion elicited in a lab-in-the-field experiment can predict field behavior (Shoji, 2020). In Bangladesh, this author showed that those with higher guilt sensitivity have a greater credit accessibility and credit-worthiness. In addition, individuals suffer from less property crime in villages

with a higher guilt-sensitivity neighbourhood. These results suggest that guilt aversion measured in the lab can indeed explain field behaviors. It strengthens our confidence in the potential policy implications of our present lab experiments. Yet, a further step would be needed, that implies measuring guilt aversion directly through field behaviors. A major challenge, though, will be to measure beliefs, as beliefs are not typically observed in the field.

Bibliography

Abeler, J., Calaki, J., Andree, K., and Basek, C. (2010). The power of apology. *Economics Letters*, 107(2):233–235.

Balafoutas, L. (2011). Public beliefs and corruption in a repeated psychological game. *Journal of Economic Behavior & Organization*, 78(1-2):51–59.

Baumeister, R. F., Stillwell, A. M., and Heatherton, T. F. (1994). Guilt: an interpersonal approach. *Psychological bulletin*, 115(2):243.

Bellemare, C., Sebald, A., and Strobel, M. (2011). Measuring the willingness to pay to avoid guilt: estimation using equilibrium and stated belief models. *Journal of Applied Econometrics*, 26(3):437–453.

Belot, M., Bhaskar, V., and Van De Ven, J. (2012). Can observers predict trustworthiness? *Review of Economics and Statistics*, 94(1):246–259.

Grether, D. M. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly journal of economics*, 95(3):537–557.

Grossman, Z. and Van Der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association*, 15(1):173–217.

Ketelaar, T. and Tung Au, W. (2003). The effects of feelings of guilt on the behaviour of uncooperative individuals in repeated social bargaining games: An affect-as-information interpretation of the role of emotion in social interaction. *Cognition and emotion*, 17(3):429–453.

Morell, A. (2019). The short arm of guilt–an experiment on group identity and guilt aversion. *Journal of Economic Behavior & Organization*, 166:332–345.

Reinikka, R. and Svensson, J. (2004). Local capture: evidence from a central government transfer program in uganda. *The quarterly journal of economics*, 119(2):679–705.

Shoji, M. (2020). Guilt and antisocial conformism: Experimental evidence from bangladesh. Technical report, University Library of Munich, Germany.

Bibliography

Abbink, K. and Serra, D. (2012). Chapter 4 anticorruption policies: Lessons from the lab. In *New advances in experimental research on corruption*, pages 77–115. Emerald Group Publishing Limited.

Abeler, J., Calaki, J., Andree, K., and Basek, C. (2010). The power of apology. *Economics Letters*, 107(2):233–235.

Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, 87(4):1115–1153.

Amdur, D. and Schmick, E. (2013). Does the direct-response method induce guilt aversion in a trust game? *Economics Bulletin*, 33(1):687–693.

Andreoni, J., Harbaugh, W. T., and Vesterlund, L. (2010). Altruism in experiments. In *Behavioural and experimental economics*, pages 6–13. Springer.

Andreoni, J. and Sanchez, A. (2020). Fooling myself or fooling observers? avoiding social pressures by manipulating perceptions of deservingness of others. *Economic Inquiry*, 58(1):12–33.

Andrighetto, G., Grieco, D., and Tummolini, L. (2015). Perceived legitimacy of normative expectations motivates compliance with social norms when nobody is watching. *Frontiers in Psychology*, 6:1413.

Armantier, O. and Boly, A. (2013). Comparing corruption in the laboratory and in the field in burkina faso and in canada. *The Economic Journal*, 123(573):1168–1187.

Ashton, M. C. and Lee, K. (2008). The prediction of honesty–humility-related criteria by the hexaco and five-factor models of personality. *Journal of Research in Personality*, 42(5):1216–1228.

Attanasi, G., Battigalli, P., and Manzoni, E. (2016). Incomplete-information models of guilt aversion in the trust game. *Management Science*, 62(3):648–667.

Attanasi, G., Battigalli, P., Manzoni, E., and Nagel, R. (2018). Belief-dependent preferences and reputation: Experimental analysis of a repeated trust game. *Journal of Economic Behavior & Organization*.

Attanasi, G., Battigalli, P., Manzoni, E., and Nagel, R. (2019a). Belief-dependent preferences and reputation: Experimental analysis of a repeated trust game. *Journal of Economic Behavior & Organization*, 167:341–360.

Attanasi, G., Battigalli, P., and Nagel, R. (2013a). Disclosure of belief-dependent preferences in a trust game. Technical report, No. 506, IGIER (Innocenzo Gasparini Institute for Economic Research), Bocconi University.

Attanasi, G., Battigalli, P., Nagel, R., et al. (2010). Disclosure of belief-dependent preferences in the trust game. In *BQGT*, pages 51–1.

Attanasi, G., Battigalli, P., Nagel, R., et al. (2013b). Disclosure of belief-dependent preferences in the trust game. *IGIER Working Papers*, 506.

Attanasi, G., Rimbaud, C., and Villeval, M. C. (2019b). Embezzlement and guilt aversion. *Journal of Economic Behavior & Organization*, 167:409–429.

Azar, O. H. (2019). The influence of psychological game theory. *Journal of Economic Behavior & Organization*, 167:445–453.

Bacharach, M., Guerra, G., and Zizzo, D. J. (2007). The self-fulfilling property of trust: An experimental study. *Theory and Decision*, 63(4):349–388.

Bahr, G. and Requate, T. (2014). Reciprocity and giving in a consecutive three-person dictator game with social interaction. *German Economic Review*, 15(3):374–392.

Balafoutas, L. (2011). Public beliefs and corruption in a repeated psychological game. *Journal of Economic Behavior & Organization*, 78(1-2):51–59.

Balafoutas, L. and Fornwagner, H. (2017). The limits of guilt. *Journal of the Economic Science Association*, 3(2):137–148.

Balafoutas, L. and Sutter, M. (2017). On the nature of guilt aversion: Insights from a new methodology in the dictator game. *Journal of Behavioral and Experimental Finance*, 13:9–15.

Barr, A., Lindelow, M., and Serneels, P. (2009). Corruption in public service delivery: An experimental analysis. *Journal of Economic Behavior & Organization*, 72(1):225–239.

Battigalli, P., Charness, G., and Dufwenberg, M. (2013). Deception: The role of guilt. *Journal of Economic Behavior & Organization*, 93:227–232.

Battigalli, P., Corrao, R., and Dufwenberg, M. (2019a). Incorporating belief-dependent motivation in games. *Journal of Economic Behavior & Organization*, 185-218:185–218.

Battigalli, P., Corrao, R., and Sanna, F. (2019b). Epistemic game theory without types structures: An application to psychological games. *IGIER Working Papers*, 641.

Battigalli, P. and Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, 97(2):170–176.

Battigalli, P. and Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory*, 144(1):1–35.

Battigalli, P. and Dufwenberg, M. (2020). Belief-dependent motivations and psychological game theory.

Battigalli, P., Dufwenberg, M., and Smith, A. (2019c). Frustration, aggression, and anger in leader-follower games. *Games and Economic Behavior*, 117:15–39.

Baumeister, R. F., Reis, H. T., and Delespaul, P. A. (1995). Subjective and experiential correlates of guilt in daily life. *Personality and Social Psychology Bulletin*, 21(12):1256–1268.

Baumeister, R. F., Stillwell, A. M., and Heatherton, T. F. (1994). Guilt: an interpersonal approach. *Psychological bulletin*, 115(2):243.

Beck, A., Kerschbamer, R., Qiu, J., and Sutter, M. (2013). Shaping beliefs in experimental markets for expert services: Guilt aversion and the impact of promises and money-burning options. *Games and Economic Behavior*, 81:145–164.

Becker, G. S. and Stigler, G. J. (1974). Law enforcement, malfeasance, and compensation of enforcers. *The Journal of Legal Studies*, 3(1):1–18.

Beekman, G., Bulte, E., and Nillesen, E. (2014). Corruption, investments and contributions to public goods: Experimental evidence from rural liberia. *Journal of public economics*, 115:37–47.

Bell, D. E. (1985). Reply—putting a premium on regret. *Management Science*, 31(1):117–122.

Bellemare, C., Kröger, S., and Van Soest, A. (2008). Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities. *Econometrica*, 76(4):815–839.

Bellemare, C., Sebald, A., and Strobel, M. (2011). Measuring the willingness to pay to avoid guilt: estimation using equilibrium and stated belief models. *Journal of Applied Econometrics*, 26(3):437–453.

Bellemare, C., Sebald, A., and Suetens, S. (2017). A note on testing guilt aversion. *Games and Economic Behavior*, 102:233–239.

Bellemare, C., Sebald, A., and Suetens, S. (2018). Heterogeneous guilt sensitivities and incentive effects. *Experimental Economics*, 21(2):316–336.

Bellemare, C., Sebald, A., and Suetens, S. (2019). Guilt aversion in economics and psychology. *Journal of Economic Psychology*, 73:52–59.

Belot, M., Bhaskar, V., and Van De Ven, J. (2012). Can observers predict trustworthiness? *Review of Economics and Statistics*, 94(1):246–259.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American economic review*, 96(5):1652–1678.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122–142.

Bock, O., Baetge, I., and Nicklisch, A. (2014). hroot: Hamburg registration and organization online tool. *European Economic Review*, 71:117–120.

Bodner, R. and Prelec, D. (2003). Self-signaling and diagnostic utility in everyday decision making. *The psychology of economic decisions*, 1(105):26.

Bolton, G. E. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American economic review*, 90(1):166–193.

Boly, A., Gillanders, R., and Miettinen, T. (2016). Deterrence, peer effect, and legitimacy in anti-corruption policy-making: An experimental analysis. *WIDER Working Paper*.

Bracht, J. and Regner, T. (2013). Moral emotions and partnership. *Journal of Economic Psychology*, 39:313–326.

Buskens, V. and Raub, W. (2013). *Rational choice research on social dilemmas: embeddedness effects on trust*. Russell Sage: New York, NY, USA.

Canagarajah, S. and Ye, X. (2001). *Public health and education spending in Ghana in 1992–98: Issues of equity and efficiency*. World Bank Publications.

Caplin, A. and Leahy, J. (2004). The social discount rate. *Journal of political Economy*, 112(6):1257–1268.

Cartwright, E. (2019a). Guilt aversion and reciprocity in the performance-enhancing drug game. *Journal of Sports Economics*, 20(4):535–555.

Cartwright, E. (2019b). A survey of belief-based guilt aversion in trust and dictator games. *Journal of Economic Behavior & Organization*, 167:430–444.

Chang, L. J., Smith, A., Dufwenberg, M., and Sanfey, A. G. (2011). Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron*, 70(3):560–572.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6):1579–1601.

Charness, G. and Dufwenberg, M. (2011). Participation. *American Economic Review*, 101(4):1211–37.

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

Chen, S., Heese, C., et al. (2020). Motivated information acquisition in social decisions. Technical report, University of Bonn and University of Mannheim, Germany.

Chlaß, N., Gangadharan, L., and Jones, K. (2015). Charitable giving and intermediation. *Jena Economic Research Papers*.

Ciccarone, G., Di Bartolomeo, G., and Papa, S. (2020). The rationale of in-group favoritism: An experimental test of three explanations. *Games and Economic Behavior*, 124:554–568.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Cohen, T. R., Wolf, S. T., Panter, A. T., and Insko, C. A. (2011). Introducing the gasp scale: a new measure of guilt and shame proneness. *Journal of personality and social psychology*, 100(5):947.

Conconi, P., DeRemer, D. R., Kirchsteiger, G., Trimarchi, L., and Zanardi, M. (2017). Suspiciously timed trade disputes. *Journal of International Economics*, 105:57–76.

Cooper, D. (2009). Other regarding preferences: a survey of experimental results in j. kagel & a. roth. *The handbook of experimental economics*, 2.

Cox, J. C., Kerschbamer, R., and Neururer, D. (2016). What is trustworthiness and what drives it? *Games and Economic Behavior*, 98:197–218.

d'Adda, G., Drouvelis, M., and Nosenzo, D. (2016). Norm elicitation in within-subject designs: Testing for order effects. *Journal of Behavioral and Experimental Economics*, 62:1–7.

d'Adda, G., Gao, Y., Golman, R., and Tavoni, M. (2018). It's so hot in here: Information avoidance, moral wiggle room, and high air conditioning usage.

Dai, Z., Galeotti, F., and Villeval, M. C. (2018). Cheating in the lab predicts fraud in the field: An experiment in public transportation. *Management Science*, 64(3):1081–1100.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.

Danilov, A., Khalmetski, K., and Sliwka, D. (2019). Descriptive norms and guilt aversion. *Mimeo*.

DellaVigna, S., Lindner, A., Reizer, B., and Schmieder, J. F. (2017). Reference-dependent job search: Evidence from hungary. *The Quarterly Journal of Economics*, 132(4):1969–2018.

Dhami, S., Wei, M., and al Nowaihi, A. (2017). Public goods games and psychological utility: Theory and evidence. *Journal of Economic Behavior & Organization*.

Dhami, S., Wei, M., and al Nowaihi, A. (2019). Public goods games and psychological utility: Theory and evidence. *Journal of Economic Behavior & Organization*, 167:361–390.

Di Bartolomeo, G., Dufwenberg, M., Papa, S., and Passarelli, F. (2018). Promises, expectations & causation. *Games and Economic Behavior*.

Di Bartolomeo, G., Dufwenberg, M., Papa, S., and Passarelli, F. (2019). Promises, expectations & causation. *Games and Economic Behavior*, 113:137–146.

Di Falco, S., Magdalou, B., Masclet, D., Villeval, M. C., and Willinger, M. (2016). Can transparency of information reduce embezzlement? experimental evidence from tanzania. *GATE Working Papers*, (1618).

Di Tella, R., Perez-Truglia, R., Babino, A., and Sigman, M. (2015). Conveniently upset: Avoiding altruism by distorting beliefs about others' altruism. *American Economic Review*, 105(11):3416–42.

Di Tella, R. and Schargrodsky, E. (2003). The role of wages and auditing during a crackdown on corruption in the city of buenos aires. *The Journal of Law and Economics*, 46(1):269–292.

Ditto, P. H. and Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of personality and social psychology*, 63(4):568.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3):522–550.

Dollard, J., Miller, N. E., Doob, L. W., Mowrer, O. H., and Sears, R. R. (1939). Frustration and aggression.

Drugov, M., Hamman, J., and Serra, D. (2014). Intermediaries in corruption: an experiment. *Experimental Economics*, 17(1):78–99.

Dufwenberg, M. (2002). Marital investments, time consistency and emotions. *Journal of Economic Behavior & Organization*, 48(1):57–69.

Dufwenberg, M. and Dufwenberg, M. A. (2018). Lies in disguise–a theoretical analysis of cheating. *Journal of Economic Theory*, 175:248–264.

Dufwenberg, M., Gächter, S., and Hennig-Schmidt, H. (2011). The framing of games and the psychology of play. *Games and Economic Behavior*, 73(2):459–478.

Dufwenberg, M. and Gneezy, U. (2000). Measuring beliefs in an experimental lost wallet game. *Games and economic Behavior*, 30(2):163–182.

Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and economic behavior*, 47(2):268–298.

Dufwenberg, M. and Kirchsteiger, G. (2019). Modelling kindness. *Journal of Economic Behavior & Organization*, 167:228–234.

Dufwenberg, M. and Nordblom, K. (2018). Tax evasion with a conscience.

Dufwenberg, M. and Patel, A. (2019). Introduction to special issue on psychological game theory. *Journal of Economic Behavior & Organization*, 167(C)(3):181–184.

Easterlin, R. A. (1995). Will raising the incomes of all increase the happiness of all? *Journal of Economic Behavior & Organization*, 27(1):35–47.

Ederer, F. and Stremitzer, A. (2017). Promises and expectations. *Games and Economic Behavior*, 106:161–178.

Ellingsen, T., Johannesson, M., Tjøtta, S., and Torsvik, G. (2010). Testing guilt aversion. *Games and Economic Behavior*, 68(1):95–107.

Elster, J. (1998). Emotions and economic theory. *Journal of economic literature*, 36(1):47–74.

Engler, Y., Kerschbamer, R., and Page, L. (2018a). Guilt averse or reciprocal? looking at behavioral motivations in the trust game. *Journal of the Economic Science Association*, 4(1):1–14.

Engler, Y., Kerschbamer, R., and Page, L. (2018b). Why did he do that? using counterfactuals to study the effect of intentions in extensive form games. *Experimental Economics*, 21(1):1–26.

Erkut, H., Nosenzo, D., and Sefton, M. (2015). Identifying social norms using coordination games: Spectators vs. stakeholders. *Economics Letters*, 130:28–31.

Exley, C. L. (2016). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies*, 83(2):587–628.

Fan, C. S., Lin, C., and Treisman, D. (2009). Political decentralization and corruption: Evidence from around the world. *Journal of Public Economics*, 93(1-2):14–34.

Farber, H. S. (2005). Is tomorrow another day? the labor supply of new york city cabdrivers. *Journal of political Economy*, 113(1):46–82.

Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2009). Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3):817–868.

Feiler, L. (2014). Testing models of information avoidance with binary choice dictator games. *Journal of Economic Psychology*, 45:253–267.

Ferraz, C., Finan, F., and Moreira, D. B. (2012). Corrupting learning: Evidence from missing federal education funds in brazil. *Journal of Public Economics*, 96(9-10):712–726.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, 10(2):171–178.

Fischbacher, U. and Föllmi-Heusi, F. (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association*, 11(3):525–547.

Fong, C. M. and Oberholzer-Gee, F. (2011). Truth in giving: Experimental evidence on the welfare effects of informed giving to the poor. *Journal of Public Economics*, 95(5-6):436–444.

Forsythe, R., Horowitz, J. L., Savin, N. E., and Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic behavior*, 6(3):347–369.

Freddi, E. (2019). Do people avoid morally relevant information? evidence from the refugee crisis. *Review of Economics and Statistics*, pages 1–45.

Friedrichsen, J., Momsen, K., Piasenti, S., et al. (2020). Ignorance, intention and stochastic outcomes. Technical report.

Garcia, T., Massoni, S., and Villeval, M. C. (2020). Ambiguity and excuse-driven behavior in charitable giving. *European Economic Review*, 124:103412.

Geanakoplos, J., Pearce, D., and Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and economic Behavior*, 1(1):60–79.

Ghidoni, R. and Ploner, M. (2020). When do the expectations of others matter? experimental evidence on distributional justice and guilt aversion. *Theory and Decision*, pages 1–46.

Gill, D. and Prowse, V. (2012). A structural analysis of disappointment aversion in a real effort competition. *American Economic Review*, 102(1):469–503.

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, 95(1):384–394.

Golman, R., Hagmann, D., and Loewenstein, G. (2017). Information avoidance. *Journal of Economic Literature*, 55(1):96–135.

Grether, D. M. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly journal of economics*, 95(3):537–557.

Grossman, Z. and Van Der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association*, 15(1):173–217.

Grubiak, K. et al. (2019). Exploring image motivation in promise keeping-an experimental investigation. Technical report, School of Economics, University of East Anglia, Norwich, UK.

Guala, F., Mittone, L., and Ploner, M. (2013). Group membership, team preferences, and expectations. *Journal of Economic Behavior & Organization*, 86:183–190.

Guerra, G. and Zizzo, D. J. (2004). Trust responsiveness and beliefs. *Journal of Economic Behavior & Organization*, 55(1):25–30.

Güth, W., Ploner, M., and Regner, T. (2009). Determinants of in-group bias: Is group affiliation mediated by guilt-aversion? *Journal of Economic Psychology*, 30(5):814–827.

Haisley, E. C. and Weber, R. A. (2010). Self-serving interpretations of ambiguity in other-regarding behavior. *Games and economic behavior*, 68(2):614–625.

Hauge, K. E. (2016). Generosity and guilt: The role of beliefs and moral standards of others. *Journal of Economic Psychology*, 54:35–43.

Heidhues, P. and Kőszegi, B. (2008). Competition and price variation when consumers are loss averse. *American Economic Review*, 98(4):1245–68.

Inderst, R., Khalmetski, K., and Ockenfels, A. (2019). Sharing guilt: How better access to information may backfire. *Management Science*, 65(7):3322–3336.

Ismayilov, H. and Potters, J. (2016). Why do promises affect trustworthiness, or do they? *Experimental Economics*, 19(2):382–393.

Kajackaite, A. (2015). If i close my eyes, nobody will get hurt: The effect of ignorance on performance in a real-effort experiment. *Journal of Economic Behavior & Organization*, 116:518–524.

Kajackaite, A. and Gneezy, U. (2017). Incentives and cheating. *Games and Economic Behavior*, 102:433–444.

Kawagoe, T. and Narita, Y. (2014). Guilt aversion revisited: An experimental test of a new model. *Journal of Economic Behavior & Organization*, 102:1–9.

Ketelaar, T. and Tung Au, W. (2003). The effects of feelings of guilt on the behaviour of uncooperative individuals in repeated social bargaining games: An affect-as-information interpretation of the role of emotion in social interaction. *Cognition and emotion*, 17(3):429–453.

Khalmetski, K. (2016). Testing guilt aversion with an exogenous shift in beliefs. *Games and Economic Behavior*, 97:110–119.

Khalmetski, K., Ockenfels, A., and Werner, P. (2015). Surprising gifts: Theory and laboratory evidence. *Journal of Economic Theory*, 159:163–208.

Knetsch, J. L. and Wong, W.-K. (2009). The endowment effect and the reference state: Evidence and manipulations. *Journal of Economic Behavior & Organization*, 71(2):407–413.

Köbis, N. C., van Prooijen, J.-W., Righetti, F., and Van Lange, P. A. (2016). Prospection in individual and interpersonal corruption dilemmas. *Review of General Psychology*, 20(1):71–85.

Kőszegi, B. and Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, 121(4):1133–1165.

Kőszegi, B. and Rabin, M. (2007). Reference-dependent risk attitudes. *American Economic Review*, 97(4):1047–1073.

Krupka, E. L., Leider, S., and Jiang, M. (2017). A meeting of the minds: informal agreements and social norms. *Management Science*, 63(6):1708–1729.

Krupka, E. L. and Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3):495–524.

Larson, T. and Capra, C. M. (2009). Exploiting moral wiggle room: Illusory preference for fairness? a comment. *Judgment and decision Making*, 4(6):467.

Le Quement, M. T., Patel, A., et al. (2018). Communication as gift-exchange. Technical report, School of Economics, University of East Anglia, Norwich, UK.

Loewenstein, G. and Molnar, A. (2018). The renaissance of belief-based utility in economics. *Nature Human Behaviour*, 2(3):166–167.

Loomes, G. and Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The economic journal*, 92(368):805–824.

Mannahan, R. (2019). Self-esteem and rational self-handicapping. *Unpub'lished*.

Massi Lindsey, L. L. (2005). Anticipated guilt as behavioral motivation: An examination of appeals to help unknown others through bone marrow donation. *Human Communication Research*, 31(4):453–481.

Mazar, N., Amir, O., and Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of marketing research*, 45(6):633–644.

Miettinen, T. and Suetens, S. (2008). Communication and guilt in a prisoner's dilemma. *Journal of Conflict Resolution*, 52(6):945–960.

Morell, A. (2019). The short arm of guilt–an experiment on group identity and guilt aversion. *Journal of Economic Behavior & Organization*, 166:332–345.

Motro, D., Ordóñez, L. D., Pittarello, A., and Welsh, D. T. (2018). Investigating the effects of anger and guilt on unethical behavior: A dual-process approach. *Journal of Business Ethics*, 152(1):133–148.

Moulton, R. W., Burnstein, E., Liberty Jr, P. G., and Altucher, N. (1966). Patterning of parental affection and disciplinary dominance as a determinant of guilt and sex typing. *Journal of Personality and Social Psychology*, 4(4):356.

Murphy, R. O., Ackermann, K. A., and Handgraaf, M. (2011). Measuring social value orientation. *Judgment and Decision making*, 6(8):771–781.

Nyborg, K. (2018). Social norms and the environment. *Annual Review of Resource Economics*.

Ockenfels, A. and Werner, P. (2014a). Beliefs and ingroup favoritism. *Journal of Economic Behavior & Organization*, 108:453–462.

Ockenfels, A. and Werner, P. (2014b). Scale manipulation in dictator games. *Journal of Economic Behavior & Organization*, 97:138–142.

O'Donoghue, T. and Sprenger, C. (2018). Reference-dependent preferences. In *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 1, pages 1–77. Elsevier.

Olken, B. A. (2007). Monitoring corruption: evidence from a field experiment in indonesia. *Journal of political Economy*, 115(2):200–249.

Olken, B. A. and Pande, R. (2012). *Accessed corruption in developing countries*. MIT Annual Reviews.

Olthof, T. (2012). Anticipated feelings of guilt and shame as predictors of early adolescents' antisocial and prosocial interpersonal behaviour. *European Journal of Developmental Psychology*, 9(3):371–388.

Patel, A. and Smith, A. (2019). Guilt and participation. *Journal of Economic Behavior & Organization*, 167:279–295.

Paternoster, R., Brame, R., Mazerolle, P., and Piquero, A. (1998). Using the correct statistical test for the equality of regression coefficients. *Criminology*, 36(4):859–866.

Peeters, R. and Vorsatz, M. (2018). Simple guilt and cooperation. *University of Otago Economics Discussion Papers*, 1801.

Peeters, R. and Vorsatz, M. (2021). Simple guilt and cooperation. *Journal of Economic Psychology*, 82:102347.

Pelligra, V. (2011). Empathy, guilt-aversion, and patterns of reciprocity. *Journal of Neuroscience, Psychology, and Economics*, 4(3):161.

Pelligra, V., Reggiani, T., and Zizzo, D. J. (2020). Responding to (un) reasonable requests by an authority. *Theory and Decision*, pages 1–25.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American economic review*, pages 1281–1302.

Rabin, M. (1995). Moral preferences, moral constraints, and self-serving biases.

Regner, T. (2018). Reciprocity under moral wiggle room: Is it a preference or a constraint? *Experimental Economics*, 21(4):779–792.

Regner, T. and Harth, N. S. (2014). Testing belief-dependent models. *Jena Economic Research Papers*.

Regner, T. and Matthey, A. (2017). Actions and the self: I give, therefore i am? *Jena Economic Research Papers*, 2017:018.

Reinikka, R. and Svensson, J. (2004). Local capture: evidence from a central government transfer program in uganda. *The quarterly journal of economics*, 119(2):679–705.

Reinikka, R. and Svensson, J. (2011). The power of information in public services: Evidence from education in uganda. *Journal of Public Economics*, 95(7-8):956–966.

Reuben, E., Sapienza, P., and Zingales, L. (2009). Is mistrust self-fulfilling? *Economics Letters*, 104(2):89–91.

Ross, L., Greene, D., and House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology*, 13(3):279–301.

Rushton, J. P., Chrisjohn, R. D., and Fekken, G. C. (1981). The altruistic personality and the self-report altruism scale. *Personality and individual differences*, 2(4):293–302.

Serra-Garcia, M. and Szech, N. (2019). The (in) elasticity of moral ignorance.

Shalvi, S., Soraperra, I., van der Weele, J. J., and Villeval, M.-C. (2019). Shooting the messenger? supply and demand in markets for willful ignorance. Technical report, Tinbergen Institute Discussion Paper.

Shoji, M. (2020). Guilt and antisocial conformism: Experimental evidence from bangladesh. Technical report, University Library of Munich, Germany.

Smith, M. K., Trivers, R., and von Hippel, W. (2017). Self-deception facilitates interpersonal persuasion. *Journal of Economic Psychology*, 63:93–101.

Solda, A., Ke, C., Page, L., and Von Hippel, W. (2019). Strategically delusional. *Experimental Economics*, pages 1–28.

Spiekermann, K. and Weiss, A. (2016). Objective and subjective compliance: A norm-based explanation of 'moral wiggle room'. *Games and Economic Behavior*, 96:170–183.

Steenhaut, S. and Van Kenhove, P. (2006). The mediating role of anticipated guilt in consumers' ethical decision-making. *Journal of business ethics*, 69(3):269–288.

Tangney, J. and Fisher, K. (1995). *Self-conscious emotions: the psychology of shame, guilt and pride*. New York: The Guilford Press.

Tangney, J. P., Dearing, R. L., Wagner, P. E., and Gramzow, R. (1989). Test of self-conscious affect–3.

Van der Weele, J. J., Kulisa, J., Kosfeld, M., and Friebel, G. (2014). Resisting moral wiggle room: how robust is reciprocal behavior? *American economic Journal: microeconomics*, 6(3):256–64.

Vanberg, C. (2008). Why do people keep their promises? an experimental test of two explanations 1. *Econometrica*, 76(6):1467–1480.

Vischer, T., Dohmen, T., Falk, A., Huffman, D., Schupp, J., Sunde, U., and Wagner, G. G. (2013). Validating an ultra-short survey measure of patience. *Economics Letters*, 120(2):142–145.

Wang, X. (2011). The role of anticipated guilt in intentions to register as organ donors and to discuss organ donation with family. *Health communication*, 26(8):683–690.

Wang, X. and McClung, S. R. (2012). The immorality of illegal downloading: The role of anticipated guilt and general emotions. *Computers in Human Behavior*, 28(1):153–159.

Woods, D. and Servátka, M. (2016). Testing psychological forward induction and the updating of beliefs in the lost wallet game. *Journal of Economic Psychology*, 56:116–125.

Xiao, E. and Bicchieri, C. (2012). Words or deeds? choosing what to know about others. *Synthese*, 187(1):49–63.

Yu, H., Shen, B., Yin, Y., Blue, P. R., and Chang, L. J. (2015). Dissociating guilt-and inequity-aversion in cooperation and norm compliance. *Journal of Neuroscience*, 35(24):8973–8975.

THREE ESSAYS ON GUILT AVERSION: THEORY AND EXPERIMENTS

Claire Rimbaud

Abstract

The aim of this thesis was to question the scope of influence of guilt-aversion by investigating (i) the direction of guilt: can individuals be guilt averse toward someone who is not affected monetarily by their decisions, or only toward someone whose earnings are affected? (ii) some necessary conditions for the emergence of guilt aversion: does the vulnerability of the person to whom individuals may feel guilty impacts the emergence of guilt? (iii) the robustness of guilt aversion against self-serving bias: are individuals strategic in their information acquisition about others' expectations in order to avoid triggering their guilt?

On the one hand, our work revealed for the very first time that people can be guilt-averse even toward people who are not affected monetarily by their decisions. This finding, obtained in the context of an embezzlement game (Chapter 1), was further extended to new contexts where guilt aversion was systematically observed toward players independently of their vulnerability (Chapter 2). On the other hand, although guilt aversion appeared to generalize to a variety of situations, we demonstrated that its robustness may be challenged in situations where the decision-makers have the possibility to avoid the tension between their monetary incentives and their belief-dependent concerns (Chapter 3).

Keywords: Guilt Aversion; Psychological Game Theory; Experiment; Corruption; Vulnerability; Information Acquisition

Résumé

L'objectif de cette thèse était de questionner la sphère d'influence de l'aversion à la culpabilité en étudiant (i) la direction de la culpabilité : les individus peuvent-ils avoir une aversion à la culpabilité envers une personne qui n'est pas affectée monétairement par leurs décisions, ou seulement envers une personne dont les revenus sont affectés ? (ii) certaines conditions nécessaires à l'émergence de l'aversion à la culpabilité : la vulnérabilité de la personne envers laquelle les individus peuvent se sentir coupables a-t-elle un impact sur l'émergence de la culpabilité ? (iii) la robustesse de l'aversion à la culpabilité face aux biais égoïstes : les individus sont-ils stratégiques dans leur acquisition d'informations sur les attentes des autres afin d'éviter de déclencher leur culpabilité ?

D'une part, notre travail a révélé pour la première fois que les gens peuvent avoir une aversion à la culpabilité même envers des personnes qui ne sont pas affectées financièrement par leurs décisions. Cette découverte, obtenue dans le contexte d'un jeu de détournement de fonds (Chapitre 1), a été étendue à de nouveaux contextes où l'aversion à la culpabilité a été systématiquement observée envers des joueurs indépendamment de leur vulnérabilité (Chapitre 2). D'autre part, bien que l'aversion à la culpabilité semble se généraliser à une diversité de situations, nous avons démontré que sa robustesse peut être remise en cause dans des situations où les décideurs ont la possibilité d'éviter la tension entre leurs incitations monétaires et leurs préoccupations liées aux croyances (Chapitre 3).

Mots Clés: Aversion à la culpabilité; Théorie des jeux psychologique; Expérience; Corruption; Vulnérabilité; Acquisition d'informations