Overconfidence as an interpersonal strategy

Alice Solda

N° d'ordre NNT : 2020LYSE2010

THESE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de

L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 486 Sciences Économique et de Gestion

Discipline : Sciences économiques

Soutenue publiquement le 7 février 2020, par :

Alice SOLDA

La surconfiance en soi :

une stratégie interpersonnelle.

Devant le jury composé de :

Béatrice BOULU-RESHEF, Professeure des universités, Université d'Orléans, Présidente

Guillaume HOLLARD, Professeur des universités, CREST École Polytechnique, Palaiseau, Rapporteur

Christiane SCHWIEREN, Professeure, Ruprecht-Karls-Universität Heidelberg, Rapporteure

Benoît TARROUX, Professeur des universités, Gate Lyon Saint-Etienne, Examinateur

Marc WILLINGER, Professeur des universités, Université Montpellier, Examinateur

Marie-Claire VILLEVAL, Directrice de Recherche, C.N.R.S., Co-Directrice de thèse

Lionel PAGE, Professeur d'université, University of Technology Sidney, Co-Directeur de Thèse

Changxia, KE, Docteur, QUT, Co-Directrice de thèse

Contrat de diffusion

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

PHD THESIS (Economics)

Overconfidence as an Interpersonal Strategy

By

Alice Soldà

Principal Supervisors:
Dr. Changxia Ke (QUT)
Prof. Lionel Page (UTS)
Prof. Marie Claire Villeval (GATE)

Associate Supervisors:
Dr. Gregory kubitz (QUT)
Prof. William von Hippel (UQ)

Committee:
Prof. Béatrice Boulu-Reshef (Université d'Orléans)
Prof. Guillaume Hollard (Polytechnique) - Reviewer
Prof. Christiane Schwieren (Heidelberg University) - Reviewer
Prof. Benoît Tarroux (GATE)
Prof. Marc Willinger (Université de Montpellier)

November 18, 2019

Acknowledgments

I would have never gone that far without the precious help, support and love of many special people. My gratitude, affection and respect goes first and foremost to Marie Claire, who believed in me when I was just an undergraduate student. You taught me that seeking greatness takes a lot of time, even more sweat and sometimes tears (not often blood, fortunately) but that efforts and determination are always rewarded. You are an inspiration, an incredible mentor and the most patient co-author a young academic can hope for. I learned so much by your side and if I can become half of the academic you are, I will consider my life full-filled.

I am also deeply indebted to my supervisors in Australia without whom this project would have never seen the light. You undertook the challenging task to make an academic out of me in such a short amount of time and I hope I rose up to your expectations. I have still so much to learn from all of you. Thank you Lionel, for giving me the opportunity to be part of this adventure. Debating and exchanging ideas with you was as exciting as it was challenging. Changxia, your sharp eye avoided me lots of moments of embarrassment. I have grown so much as a researcher, as well as a person, by your side. Bill, I am going to miss your stories and your incredible wisdom. Thank you for opening my eyes to the bigger picture. And finally Greg, even though our paths crossed at the end of this venture, thank you for your writing tips and precious help with my job applications.

Running experiments would not be as fun without going through the hassle of programming them. Anthony, Gaurav, Yi and Quentin, I am forever in your debt for helping me to grasp the mysterious (and sometimes impenetrable) ways of Python.

A huge thanks to my QUT acquaintances, who made Brisbane feel like home. Anthony and Ambroise, who took me under their wings. Tatie and Diego who shared their home with me and introduced me to *Mystery Diners*. Laura, for understanding my soft spot for Korean dramas and pop culture. I enjoyed every trip, movie night and bubble tea we had together. Sylvain, Florian and Ammar, who took me on memorable trips, from Wara Wara to the white sandy beaches of Fraser Island. Jeremy, Martin, Richard and Caleb, who fed me with fried chicken and made sure I never get bored after 4.30pm.

On the other side of the world, this adventure would have not been the same without my colleagues at GATE: Rémi, Vincent, Julien, Clément, Maxime, Thomas and Morgan. A special thanks to Charlotte and Claire, who always have the right

words to help me get through rough times. I also thank my dearest friends, Sy, Lise, Nice and Alexis, that I have known my whole life for their unconditional love. It would take another thesis to tell how lucky I am to have you. I am deeply grateful to my parents, for raising me to be a person I can be proud of; and to my partner, Victor, for his unconditional support - on the rift and IRL - I am thrilled to start this new journey with you by my side.

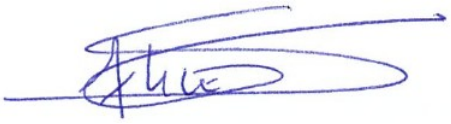
And because the path to the *TOP* was not always bright, I am forever thankful to you, who shines the most. Your voice *is* the light who kept me going through *the* darkest nights. Being able to express my gratitude through something I put the *best* of myself into means the world to me.

I dedicate this thesis to my grandmother who was my best friend for the twenty first years of my life. I miss you terribly.

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Brisbane, 18/11/2019



Abstract

Standard economic models assume that individuals collect and process information in a way that gives them a relatively accurate perception of reality. However, this assumption is often violated. Data shows that individuals often form positively biased beliefs about themselves, which can have detrimental economic consequences. This thesis aims to explain the persistence of overconfidence in social interactions by showing the existence of strategic benefits of being overconfident that offset its social cost.

Using a series of laboratory experiments, this thesis shows that (i) overconfidence emerges primarily when it provides an advantage in social interactions (Chapter 2) and (ii) identify situations in which overconfidence is likely to be socially detrimental (Chapter 3 and 4). This thesis contributes to the literature by enhancing our understanding of the situational determinants of overconfidence in social interactions and lay the foundations to improve policies intended to prevent or limit its negative effects.

Keywords: Experimental Economics; Behavioral Economics; Self-Deception; Motivated Beliefs; Overconfidence; Decision-Making; Negotiations; Leadership

Résumé détaillé en français[1]

Les individus prennent des décisions en fonction de leur perception de leurs propres capacités. Les étudiants choisissent une filière universitaire dans laquelle ils pensent pouvoir obtenir de bons résultats. Un couple peut décider de ne pas avoir d'enfant s'il s'attend à perdre son emploi dans l'avenir. Un État décide de déclarer la guerre à un autre s'il pense que son armée est la plus forte. Les conséquences de ces décisions dépendent - en partie du moins - de notre faculté à évaluer avec précision nos propres capacités (Dunning et al., 2004). La plupart des théories classiques en économie (Von Neumann and Morgenstern, 1953) et en psychologie (Maslow, 1950; Körding and Wolpert, 2004) supposent que les individus collectent et traitent l'information d'une manière qui leur donne une perception relativement précise de la réalité. Dans les modèles bayésiens, les agents actualisent leur croyance en fonction de signaux reçus sur la probabilité d'un événement, compte tenu de leur croyance initiale. L'approche bayésienne offre un moyen simple et fiable d'évaluer la façon dont l'information est traitée et comment celle-ci influence la prise de décision. Cependant, des données empiriques indiquent que cette hypothèse n'est pas toujours vérifiée. Tversky and Kahneman (1974) ont montré que les individus sont sujets à divers biais dans la mise à jour de leurs croyances sur les événements probabilistes. L'un des biais les plus documentés concernant la mise à jour des croyances est l'excès de confiance ('surconfiance').

1. La surconfiance en tant que stratégie

Moore and Healy (2008) définissent la surconfiance comme une perception positivement biaisée de soi qui consiste à "*(1) surestimer sa performance réelle, (2) surestimer sa performance par rapport à celle des autres, et (3) donner une précision excessive à ses propres croyances*". Il existe de nombreuses preuves dans la littérature que les individus sont surconfiants dans divers domaines. Les gens pensent qu'ils sont plus attirants (Epley and Whitchurch, 2008), plus intelligents (Gabriel et al., 1994), meilleurs universitaires (Cross, 1977) et meilleurs conducteurs (Svenson, 1981) qu'ils ne le sont réellement. Bien que la surconfiance puisse parfois être attribuée à des erreurs de calcul ou à l'asymétrie d'information (Chambers and Windschitl, 2004), de plus en plus de résultats scientifiques suggèrent que ce biais est motivé (Bénabou, 2015; Bénabou and Tirole, 2016).

Un raisonnement motivé se produit lorsqu'un biais conduit à une augmentation des gains espérés. Des modèles théoriques ont montré que le fait d'avoir des

[1] This chapter is an extensive summary of this thesis written in French as required by Université Lumière - Lyon 2.

croyances positivement biaisées à l'égard de soi-même augmente la motivation d'entreprendre des projets (Bénabou and Tirole, 2002), améliore le rendement de l'entreprise (Bénabou, 2012) et atténue le problème de sous-investissement des efforts associé au biais de préférence pour le présent (Hong et al., 2018). Ces prédictions sont corroborées par des résultats empiriques montrant que la surconfiance augmente la provision de l'effort et le rendement d'équipe (Vialle et al., 2011), ainsi que la performance sportive au cours du temps (Murphy et al., 2017). De plus, les PDG surconfiants sont plus susceptibles d'innover et de réussir (Galasso and Simcoe, 2011; Hirshleifer et al., 2012). Des données d'enquête indiquent que les personnes plus confiantes travaillent plus dur (Puri and Robinson, 2007). Ces modèles économiques de croyances motivées mettent l'accent sur les avantages *intrapersonnels* de former des croyances positivement biaisées.

Un autre volet de la littérature propose que la surconfiance procure des avantages *interpersonnels* parce qu'elle permet d'influence les autres de manière bénéfique dans les interactions sociales (Heifetz et al., 2007; Johnson and Fowler, 2011; von Hippel and Trivers, 2011). Il a été démontré que la confiance en soi joue un rôle important dans la sélection des leaders (Shamir et al., 1993), des partenaires romantiques (Buss, 2009; Murphy et al., 2015) et des fournisseurs de services sociaux et matériels (De Jong et al., 2006), La confiance en soi détermine également l'influence sociale : les personnes confiantes sont plus crues (Penrod and Cutler, 1995) et se voient offrir de plus grandes concessions dans les négociations (Swift and Moore, 2012).

Bien que ces résultats suggèrent qu'avoir confiance en soi procure un avantage, cela ne signifie pas nécessairement qu'être surconfiant apporte de tels avantages. Des articles scientifiques récents ont montré que c'est effectivement le cas. Murphy et al. (2017) ont suivi 894 garçons du secondaire pendant deux années scolaires et ont documenté que la surconfiance concernant les capacités sportives prédisait une popularité accrue au cours du temps. Ces avantages a être surconfiant ont été étendus à des expériences économiques en laboratoire. Charness et al. (2018) ont montré que les gens étaient plus susceptibles de reporter des niveaux de confiance plus élevés lorsqu'un excès de confiance dissuade un concurrent de participer à un tournoi. Dans l'expérience de Schwardman and van der Weele (2019), la moitié des participants s'attendent à devoir plus tard persuader leurs pairs qu'ils ont bien réussi une tâche. Les auteurs ont constaté que les participants informés de ces entretiens ont des croyances positivement biaisées au sujet de leur performance dans une tâche et que cette augmentation de confiance conduit à des gains espérés plus élevés. Dans le contexte du statut social, Anderson et al. (2012) ont constaté que les gens sont plus impressionnés par la surconfiance

que par une compétence réelle et que, par conséquent, les personnes surconfiantes obtiennent un statut social supérieur.

Ces résultats soutiennent l'idée que les personnes deviennent surconfiantes dans les interactions interpersonnelles lorsque cela augmente leur espérance de gains. Mais quels sont les méchanismes qui entrent en jeu ? Dans la section 2, je décris un cadre théorique de la psychologie évolutionniste qui propose que la surconfiance est une forme d'auto-illusion qui opère par la tromperie.

2. Se tromper soi-même pour mieux tromper les autres

L'auto-illusion

Gur and Sackeim (1979) définissent à l'origine l'auto-illusion comme un état qui exige "*deux représentations de certains aspects de la réalité, l'une exacte et l'autre systématiquement inexacte, et la partie ayant accès à l'information exacte (l'auto-illusion) doit avoir le contrôle sur l'information disponible pour l'autre partie (le soi trompé)*" (Pinker, 2011, p. 36). Cependant, il n'existe pas d'outils méthodologiques issus de l'économie ou de la psychologie permettant de montrer que les individus gardent deux versions contradictoires de la réalité. Pour cette raison, j'utiliserai la définition de von Hippel and Trivers (2011) qui assouplit cette hypothèse et propose que l'auto-illusion se produit si les individus déforment stratégiquement leur perception de la réalité. Comme le soulignent Schwardman and van der Weele (2019), cette définition "*saisit à la fois le changement stratégique et intentionnel de la confiance (la "tromperie") et le fait que les gens s'approprient ces croyances et sont prêts à parier sur elles (le "soi")*" (p. 22), ce qui met en relief le compromis entre le renforcement stratégique de soi et la précision, sur lequel porte la thèse.

L'approche évolutive

La tromperie (c.-à-d. le fait de faire croire à quelqu'un quelque chose qui est faux) est fréquemment observée dans les données empiriques (Hyman, 1989; DePaulo et al., 1996; Rosenbaum et al., 2014). Les psychologues évolutionnistes voient cette tromperie comme une stratégie dans la lutte pour des ressources rares. D'autres études ont montré que les gens mentent à ceux dont ils dépendent pour recevoir des ressources qui ne leur seraient pas fournies autrement (Steinel and De Dreu, 2004) et 50% des mensonges quotidiens visent à obtenir des ressources pour soi-même (DePaulo and Kashy, 1998). La théorie économique classique suppose qu'un agent rationnel n'est honnête que si la récompense de l'honnêteté l'emporte sur celle de la malhonnêteté (Becker, 1968). Lorsqu'il est possible de punir, les agents mentent si les bénéfices du mensonge sont suffisamment élevés

pour couvrir le coût de la punition en cas de détection (Akerlof, 1970; Bentham, 1789; Lewicki, 1983).

Cependant, un grand nombre de données économiques montrent que les individus ne mentent pas toujours, même lorsque cela est bénéfique. Sur la base de ces observations, certains modèles économiques supposent maintenant que le mensonge a des coûts intrinsèques qui entrent directement dans la fonction d'utilité (Mazar et al., 2008; Fischbacher and Föllmi-Heusi, 2013; Gneezy et al., 2018). Une explication alternative proposée par les psychologues est que le mensonge impose des charges cognitives au trompeur (Vrij and Ganis, 2014; Vrij et al., 2017). Lorsqu'il ment, le trompeur doit conserver deux versions des faits dans son esprit (la réalité et la fiction). Les premières études de Zuckerman et al. (1981) montrent que mentir peut être plus éprouvant mentalement que dire la vérité. D'autres ont montré que la tromperie est associée à une activation accrue du cortex préfrontal (Spence et al., 2008) et des processus de contrôle exécutif comme la mémoire active (Christ et al., 2008), soutenant cette idée que la tromperie impose une charge cognitive sur le trompeur.

D'une part, une tromperie réussie peut procurer des avantages substantiels au trompeur, au détriment du trompé (DePaulo, 2004). D'autre part, les individus peuvent laisser échapper des signes de tromperie, ce qui entraîne des coûts substantiels (c.-à-d. une punition) pour le trompeur si la tromperie est détectée (Boles et al., 2000; Schweitzer et al., 2006). En raison de cette constante lutte co-évolutionnaire entre le trompeur et le trompé, l'auto-illusion peut être un outil de tromperie important qui diminue la probabilité que la tromperie soit détectée (Trivers, 1976, 1985, 2000, 2010). L'idée clé de cette approche évolutive se trouve dans les premiers travaux de Robert Trivers :
"Si (comme le soutient Dawkins) la tromperie est fondamentale dans la communication animale, alors il doit y avoir une forte sélection pour repérer la tromperie et cela devrait, à son tour, sélectionner pour un certain degré d'auto-illusion, rendant certains faits et motifs inconscients afin de ne pas trahir par les signes subtils de la connaissance de soi la tromperie pratiquée. Ainsi, la vision conventionnelle selon laquelle la sélection naturelle favorise les systèmes nerveux qui produisent une vision toujours plus exacte du monde doit être une vision très naïve de l'évolution mentale."(Trivers, 1976, p. 20).

Une abondante documentation sur la détection des mensonges suggère qu'il s'agit généralement d'un exercice difficile. La méta-analyse de Bond and DePaulo (2006) rapporte un taux de détection global de 54% (à peine au-dessus du hasard). Cependant, von Hippel and Trivers (2011) et Belot and Van de Ven (2017)

soulignent l'absence d'incitations à tromper ou à détecter la tromperie (en raison des conséquences limitées ou inexistantes de la tromperie sur le trompeur et le trompé), et le contexte restreint des interactions dans les études de détection du mensonge qui peuvent expliquer le faible taux de détection documenté dans la littérature en psychologie.

Des expériences de communication plus sophistiquées ont montré une augmentation du taux de détection (DePaulo et al., 2003; Zuckerman et al., 1981; Belot and Van de Ven, 2017; von Hippel et al., 2016). Colwell et al. (2009) montrent que le fait d'avoir reçu une formation pour repérer des signes fiables de tromperie comme la dilatation de la pupille (Wang et al., 2010) ou les faux sourires (Ekman et al., 1988) conduit à un taux de détection beaucoup plus élevée (77%) que chez les observateurs non formés (57%). Mann and Vrij (2006) ont constaté un taux de détection de 72% chez les policiers lorsqu'il leur a demandé d'évaluer de réels enregistrements d'interrogatoires criminels. Les études utilisant des journaux intimes suggèrent que les gens détectent la tromperie à un taux qui est bien supérieur à celui du hasard. Les participants à l'étude de DePaulo et al. (1996) rapportent que 15 à 23% de leurs mensonges ont été détectés et 16 à 23% d'entre eux n'étaient pas certains que leurs mensonges aient été détectés. La tromperie peut également être détectée par des signes associés à la charge cognitive (Vrij, 2004; Vrij et al., 2006). Les résultats de la méta-analyse de Vrij et al. (2017) montrent qu'imposer une charge cognitive aux trompeurs conduit à une détection des mensonges de 71%. Ces résultats suggèrent que la capacité des gens à détecter la tromperie a pu être sous-estimée par le passé.

von Hippel and Trivers (2011) proposent que l'auto-illusion (i) aide à dissimuler les signes de tromperie et (ii) diminue les coûts cognitifs associés au mensonge. Cette idée a été formalisée par Aviad Heifetz et les co-auteurs qui ont écrit :
"Dans presque tous les jeux, pour presque toutes les distorsions des gains réels d'un joueur, une certaine mesure de cette distorsion est bénéfique pour le joueur en raison de l'effet résultant sur le jeu des adversaires. Par conséquent, de telles distorsions ne seront pas éliminées par un processus évolutif impliquant une dynamique de sélection monotone des gains, dans laquelle les agents dont les gains réels sont plus élevés prolifèrent au détriment des agents moins performants."(Heifetz et al., 2007, p. 2).

La perspective intrapersonnelle
L'approche psychologique standard soutient que la surconfiance est une *"stratégie coûteuse maintenue parce qu'elle apporte des avantages psychologiques aux personnes surconfiantes."* (Pinker, 2011). Cet argument a été étayé par une abon-

dante littérature empirique en psychologie qui démontre que le fait de posséder des croyances positivement biaisées sur soi-même améliore l'estime de soi (Alicke, 1985; Dunning et al., 1995) ainsi que la santé physique et mentale (Alloy and Abramson, 1979; Taylor and Brown, 1988; Taylor et al., 2000; Korn et al., 2013; Carver and Scheier, 2014). De plus, Puri and Robinson (2007) ont constaté que les personnes trop optimistes travaillent davantage, épargnent davantage, s'attendent à prendre leur retraite plus tard et sont plus susceptibles de se remarier après un divorce. Je soutiens que ces avantages psychologiques sont des produits dérivés de la fonction évolutive primaire de l'auto-illusion et n'entrent pas en conflit avec l'argument de Trivers (1976).

3. Comment se tromper soi-même ?

Nous avons vu que les gens biaisent stratégiquement leurs croyances sur eux-mêmes dans les interactions sociales lorsque cela conduit à des gains plus élevés. Mais comment ces croyances biaisées sont-elles formées et maintenues ? Dans cette section, je passe en revue les diverses stratégies auto-illusoires documentées dans la littérature. Je décris d'abord les stratégies qui surviennent avant que l'information ne soit découverte (ignorance stratégique), puis je me tourne vers les stratégies qui émergent après la découverte de l'information mais lorsque celle-ci n'est pas complètement intégrés (processus de codage biaisés). Enfin, je rapporte les stratégies qui surviennent après l'encodage de l'information (déni de réalité).

Ignorance stratégique

Les individus peuvent prévenir l'encodage d'informations indésirables dans le cerveau en évitant activement les sources d'information qui peuvent contenir de mauvaises nouvelles (Golman et al., 2017). La théorie économique classique prédit qu'il faut toujours accepter des informations gratuites qui peuvent conduire à de meilleures décisions. Cependant, il existe de nombreuses situations dans lesquelles les gens évitent activement ces informations, se privant ainsi de contributions potentiellement précieuses à la prise de décision. Les personnes à risque de problèmes de santé qui évitent des tests médicaux (parfois au détriment de leur espérance de vie) sont probablement l'illustration la plus convaincante de l'évitement actif d'informations instrumentales. En dehors de la sphère médicale, des études ont montré que les individus hésitent à se renseigner sur leur beauté et leur intelligence (Eil and Rao, 2011; Mobius et al., 2011; Burks et al., 2010), leurs aptitudes sociales (Trope et al., 2003), la valeur de leur portefeuille lorsque le marché boursier est bas (Karlsson et al., 2009; Loewenstein et al., 2016), leurs alternatives après un achat (Olson and Zanna, 1979) et même leurs gains (Huck et al., 2015). S'appuyant sur ces données empiriques, les modèles théoriques de Köszegi (2006) et Weinberg (2009) supposent que les individus considèrent non

seulement la valeur instrumentale de l'information qu'ils reçoivent, mais aussi le gain ou la perte en utilité associée à sa valence (positive ou négative). Dans ce cas, les gens sont plus susceptibles d'éviter des informations pertinentes sur eux-mêmes si leur contenu est susceptible d'altérer négativement la perception qu'ils ont d'eux-mêmes.

Lorsque l'information n'est pas gratuite (c.-à-d. que les personnes doivent consacrer du temps à la recherche de l'information ou payer pour l'obtenir), les gens peuvent favoriser des sources d'information qui correspondent à leur propre opinion (Frey, 1986). En accord avec cette idée, des études ont montré que les gens cessent de recueillir des informations lorsqu'ils apprécient les premiers retours plutôt que lorsqu'ils ne les apprécient pas (Ditto and Lopez, 1992). Les gens peuvent aussi choisir d'allouer leur attention différemment entre l'information positive et l'information négative ou menaçante. Par exemple, il a été démontré que les gens passent plus de temps à regarder des informations négatives plutôt que positives sur leur partenaire potentiel lorsqu'ils s'attendent à être rejetés (Wilson et al., 2004) et sont plus désireux de supprimer un bruit de fond sur un discours qui correspond à leur opinion que lorsqu'il ne leur correspond pas (Brock and Balloun, 1967). Les études de *eye-tracking* (suivi du regard) révèlent également que les gens accordent plus d'attention aux aspects de l'information disponible qu'ils préféreraient être vrais (Isaacowitz, 2006; Isaacowitz et al., 2008).

Processus de codage biaisés

Lorsque l'information non désirée est néanmoins codée, les gens peuvent quand même filtrer l'information négative (Taylor and Brown, 1988) et favoriser celle qu'ils sont motivés à croire (Lord et al., 1979; Dawson et al., 2006; Babcock et al., 1995a). De plus en plus de preuves empiriques supportent l'idée que les gens ont tendance à accorder plus de poids aux bonnes nouvelles qu'aux mauvaises (Eil and Rao, 2011; Mobius et al., 2011; Sharot et al., 2012; Wiswall and Zafar, 2015; Sharot and Garrett, 2016). Dans les expériences de Mobius et al. (2011) et (Eil and Rao, 2011), les auteurs ont constaté que les sujets réagissent davantage aux retours positifs sur leur performance relative et n'actualisent pas suffisamment les retours négatifs. D'autres ont montré que les gens accordent plus de poids à l'information quand elle soutient leur propre point de vue. Lorsqu'on leur présente des arguments qui soutiennent à la fois leur point de vue et le point de vue opposé d'une discussion, les personnes qui ont de solides croyances initiales d'un côté ou de l'autre ont tendance à finir par polariser encore plus la discussion (Lord et al., 1979; Dawson et al., 2002; Glaeser and Sunstein, 2013; Sunstein et al., 2016).

Déni de réalité

Même lorsque les gens sont exposés à des informations désagréables et les encodent de manière impartiale, ces informations peuvent être stratégiquement oubliées ou mal mémorisées (Crary, 1966) car elles peuvent aider à voir le comportement passé sous un jour positif (Moore, 2016). Les modèles théoriques montrent que les individus sont incités à oublier les signaux indésirables (Bénabou and Tirole, 2002) et à se rappeler les signaux négatifs avec une probabilité inférieure au pourcentage réel (Gottlieb, 2014). Ces prédictions ont été appuyées par de nombreux résultats expérimentaux. Les premiers résultats en psychology ont montré que les gens biaisent leur mémoire de leurs croyances au sujet de leurs propres compétences/caractéristiques lorsque cela les réconforte dans l'idée qu'ils se sont améliorés (Conway and Ross, 1984; Croyle et al., 2006). Les gens ont également tendance à se rappeler leur succès plutôt que leurs échecs (Korner, 1950; Mischel et al., 1976), les bonnes choses plutôt que les mauvaises (Thompson and Loewenstein, 1992; Story, 1998; Sedikides and Green, 2009), et quand ils se sont comportés de façon éthique plutôt que quand ils ne l'ont pas fait (Kouchaki and Gino, 2016).

Les données économiques tendent à montrer que les individus manipulent leurs souvenirs pour conserver une image positive d'eux-mêmes. Les individus oublient davantage leurs réponses incorrectes que les bonnes réponses de leurs performances passées dans un test de QI (Li, 2017; Chew et al., 2018; Zimmermann et al., 2019). Dans le domaine social, Shu and Gino (2012) ont constaté que les tricheurs se rappelaient moins d'articles d'un code moral que les autres. Li (2013) a constaté que les receveurs trahis dans un jeu de confiance se rappellent moins bien des resultats du jeu que les receveurs qui ont bénéficié d'une décision altruiste de leur donneur. Saucet and Villeval (2018) ont constaté que les dictateurs dans un jeu de dictateur étaient plus susceptibles de se rappeler leurs décisions altruistes que leurs décisions égoïstes.

Enfin, si l'information est correctement rappelée, les gens peuvent utiliser des stratégies d'auto-signalisation (Bénabou and Tirole, 2016) par lesquelles ils "produisent" des signaux qui leur permettent d'interpréter ultérieurement leurs choix antérieurs comme souhaitables (Quattrone and Tversky, 1984; Bodner and Prelec, 2003; Bénabou and Tirole, 2004, 2011) ou de justifier les motifs derrière ce comportement. Pour illustrer ce dernier point, Von Hippel et al. (2005) ont montré que lorsque la tricherie peut être considérée comme non intentionnelle, les personnes qui ont fait preuve d'un bais égocentrique dans un autre domaine sont plus susceptibles de tricher. Les gens sont également plus susceptibles d'éviter de s'asseoir à côté de personnes handicapées (Snyder et al., 1979) ou de venir en aide à des Afro-Américains (Saucier et al., 2005) lorsque ce comportement peut

être rationalisé.

4. Objectifs et grandes lignes

De plus en plus de preuves appuient maintenant l'idée de Trivers (1976) selon laquelle les individus se dupent eux-mêmes de manière stratégique dans les interactions sociales parce que cela conduit à des gains supérieurs. Cependant, les preuves de l'existence de l'auto-illusion et de son effet causal sont rares en économie et notre compréhension de ce phénomène reste limitée. Premièrement, peu de designs expérimentaux permettent de mesurer dans quelle mesure les gens croient ce qu'ils prétendent croire. La plupart des résultats de la littérature en psychologie reposent sur l'auto-évaluation et ne fournissent pas d'incitations monétaires directes à fournir des estimations exacte (ou un coût pour ne pas être exact). La même préoccupation se pose lorsque ces estimations sont rendues publiques.

Deuxièmement, en dépit de ces preuves, il a également été démontré que la surconfiance a des conséquences économiques indésirables. Les travailleurs surconfiants choisissent un système de paiement risqué qui conduisent à des gains plus faible qu'avec un salaire fixe inférieur (Barron and Gravert, 2018), les traders de sexe masculin surconfiants prennent des risques inconsidérés (Barber and Odean, 2001) et les propriétaires ont des attentes irréalistes concernant la valeur de leur propriété dans le futur (Case and Shiller, 2003). Ce qui est encore plus problématique, c'est que les coûts associés aux décisions prises par des personnes surconfiantes peuvent aussi affecter les autres. Comme Johnson and Fowler (2011) nous le rappellent : *Tout au long de l'histoire, on a reproché à la surconfiance d'être à l'origine de catastrophes très médiatisées comme la Première Guerre mondiale, la guerre du Vietnam, la guerre en Irak, la crise financière de 2008 et la préparation aux phénomènes environnementaux tels que l'ouragan Katrina et le changement climatique".*

En finance, Heaton (2002) et Malmendier and Tate (2005) ont constaté que les PDG trop confiants prennent de mauvaises décisions d'investissement ou de fusion. La surconfiance conduit également à une entrée excessive sur les marchés financiers (Camerer and Lovallo, 1999) et à des délais inefficaces dans le cadre des négociations (Ortner, 2013). En psychologie, Shipman and Mumford (2011) ont constaté que les leaders surconfiants sont prompts à poursuivre des objectifs personnalisés aux dépens des autres, font des erreurs et ne parviennent pas à identifier les déficiences et les problèmes. A travers les trois essais qui constituent cette thèse, j'entends aborder cette question en proposant une série d'expériences dans lesquelles (i) les croyances des participants sont manipulées de manière exogène ;

(ii) l'écart entre les croyances des participants et la performance réelle des participants est clairement mesuré ; et (iii) le fait de maintenir des croyances biaisées a un coût direct. Ces caractéristiques permettent d'identifier de manière causale l'impact individuel et social de la surconfiance.

Le chapitre 2 contribue à la littérature (i) en fournissant des preuves convaincantes que la surconfiance se manifeste dans les interactions sociales principalement lorsqu'elle procure un avantage ; et (ii) en montrant que le degré auquel les gens peuvent se duper eux-mêmes dépend de la marge de manœuvre mentale dont ils disposent pour le faire. Dans cette expérience, les participants sont encouragés soit à se forger des croyances précises sur leur performance à un test, soit à convaincre un groupe d'autres participants qu'ils ont bien réussis. Je varie également la capacité des participants à recueillir librement de l'information sur leur performance. Je montre que les participants qui s'attendent à devoir convaincre autrui sélectionnent des informations qui leur permette d'obtenir des retours plus positives sur leur performance. Les résultats montrent également que les participants qui s'attendent à devoir convaincre autrui sont plus confiants que ceux qui ne s'y attendent pas ; et cette augmentation de la confiance a un effet positif sur leur capacité à convaincre.

La deuxième partie de cette thèse porte sur deux types d'interactions dans lesquelles l'intérêt personnel entre en conflit avec l'intérêt social : les négociations bilatérales (chapitre 3) et la sélection de dirigeants (chapitre 4). Les chapitres 3 et 4 contribuent à la littérature en montrant que, dans de telles interactions, la surconfiance est susceptible d'être préjudiciable à la société, tandis que les avantages d'être surconfiant pour l'individu surconfiant sont maintenus.

Au chapitre 3, j'examine l'effet de la confiance relative sur l'issue d'une situation de négociation bilatérale. Les participants sont appariés et chaque paire effectue une tâche pour gagner des points qui sont attribués à un compte commun. A la fin de la tâche, je manipule la confiance des participants quant à leur performance par rapport à celle de leur partenaire. Le compte commun est ensuite divisé en deux parts inégales (70%-30% du compte commun) et chaque paire doit convenir de la répartition des parts entre ses membres. J'utilise un processus de négociation en trois étapes dans lequel les paires de participants ont l'occasion de parvenir à un accord à chacune de ces trois étapes. Toutefois, plus chaque paire met de temps à se mettre d'accord, plus le montant final de chaque part sera faible. Si la paire ne parvient pas à un accord, les deux membres se retrouvent les mains vides. En accord avec la littérature, j'ai constaté qu'une confiance excessive conduit à une perte d'efficience dans le processus de négociation en raison de retards coûteux

dans le temps nécessaire à la conclusion d'un accord et d'impasses. En revanche, les résultats montrent que les négociateurs qui sont relativement plus confiants que leurs partenaires sont plus susceptibles d'obtenir des gains supérieurs.

Dans le chapitre 4, j'examine si les individus deviennent stratégiquement surconfiants lorsque devenir le leader donnent accès à des ressources profitables et dans quelle mesure les décisions prises par des leaders surconfiants sont préjudiciables pour le groupe. Dans cette expérience, les participants entreprennent une tâche et sont ensuite appariés en groupes de quatre. Chaque groupe doit choisir un leader qui fera une série de choix binaires risqués au nom du groupe. Pour chaque décision, la probabilité d'obtenir le résultat désiré dépend de la probabilité que le leader soit classé comme étant le plus performant de son groupe. Ainsi, choisir le membre du groupe ayant le mieux performé à la tâche en tant que leader maximise le gain social. Avant de faire leur choix, les membres du groupe sont autorisés à communiquer entre eux par le biais d'une boîte de dialogue. La moitié des leaders se voient offrir un bonus. Les résultats montrent que les leaders du traitement avec le bonus (i) sont moins susceptibles d'être le membre de leur groupe ayant le mieux performés et (ii) font des choix surconfiants qui rapportent moins d'argent à leur groupe que les leaders qui n'ont pas reçu le bonus.

Ces résultats nous aident à mieux comprendre les déterminants situationnels de la surconfiance et expliquent pourquoi la surconfiance persiste, même lorsque cela entraîne des conséquences dramatiques. Même si la surconfiance peut être socialement coûteuse, il n'y a aucune raison de s'attendre à ce que les individus ne soient pas surconfiants quand cela reste bénéfique pour l'individu surconfiant. La plupart des interactions dans la vie réelle impliquent des situations où l'intérêt privé des agents est orthogonal aux intérêts des autres agents engagés dans l'interaction : marchés, compétitions, biens publics, etc. Par conséquent, ces connaissances devraient aider les organisations à anticiper les situations où la surconfiance est susceptible d'apparaître. Le chapitre 5 examine la portée de ces résultats, ainsi que les prolongements possibles de cette thèse, et conclut.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

"When a person cannot deceive himself the chances are against his being able to deceive other people."

Mark Twain, Autobiography of Mark Twain.

People make decisions based on their perceptions of their own abilities. Students choose their college major depending on what subject they believe they will perform well. A couple may decide not to have a baby if they expect to loose their jobs in the future. A state resolves to declare war to another if it believes his army to be stronger. The outcome of those decisions depends - in part at least - on being able to accurately evaluate our own abilities (Dunning et al., 2004). Most standard theories in economics (Von Neumann and Morgenstern, 1953) and psychology (Maslow, 1950; Körding and Wolpert, 2004) assume that people collect and process information in a way that gives them a relatively accurate perception of reality. In Bayesian-rational models, agents update their belief conditional on signals received about the likelihood of an event, given a certain prior. The Bayesian approach offers a tractable and straightforward way to assess how information is processed and affects decision making. However, empirical evidence have shown that this assumption may often not be warranted. Tversky and Kahneman (1974) showed that individuals are subjects to various biases in belief updating about probabilistic events. One of the most documented bias related to belief updating is overconfidence.

1.1 Overconfidence as a Strategy

Moore and Healy (2008) define overconfidence as a positively biased perception of oneself that consists in "(1) overestimation of one's actual performance, (2) overplacement of one's performance relative to others, and (3) excessive precision in one's beliefs". There are abundant evidence from the literature that individuals are overconfident in various domains. People believe they are more attractive (Epley and Whitchurch, 2008), smarter (Gabriel et al., 1994), better academics (Cross, 1977) and better drivers (Svenson, 1981) than they actually are. While overconfidence can sometimes be attributed to computational errors or asymmetric information (Chambers and Windschitl, 2004), a growing body of evidence suggests that this bias can be motivated (Bénabou, 2015; Bénabou and Tirole, 2016).

Motivated reasoning occurs when deviating from Bayes' rule leads to an increase in expected payoffs. Theoretical models have shown that holding positively biased beliefs about one's own self increases one's motivation to undertake challenging projects (Bénabou and Tirole, 2002), enhances corporate performance (Bénabou, 2012) and alleviates the under-investment of effort problem associated with present bias[1] (Hong et al., 2018). These predictions are supported by empirical findings showing that overconfidence increases the provision of effort and team output of workers (Vialle et al., 2011) and sporting performance over time (Murphy et al., 2017). In addition, overconfident CEOs are more likely to innovate and be successful (Galasso and Simcoe, 2011; Hirshleifer et al., 2012). Survey data also point out that more confident individuals work harder (Puri and Robinson, 2007). These economic models of motivated beliefs focus on the *intrapersonal* advantages of holding positively biased beliefs.

Another strand of the literature proposes instead that overconfidence provides interpersonal benefits because it advantageously influences others in social interactions (Heifetz et al., 2007; Johnson and Fowler, 2011; von Hippel and Trivers, 2011). Signalling self-confidence has been shown to play an important role in the selection of leaders (Shamir et al., 1993), romantic partners (Buss, 2009; Murphy et al., 2015) and providers of social and material services (De Jong et al., 2006). Confidence also determine social influence since confident individuals are more believed (Penrod and Cutler, 1995) and are offered larger concessions in negotiations (Swift and Moore, 2012).

[1] Over-valuing early returns relative to returns that will happen in the future.

While these findings suggest that signaling confidence provides an advantage, it does not necessarily mean that being overconfident will come up with such benefits. Recent papers have shown that it is actually the case. Murphy et al. (2017) tracked 894 high school boys across two school years and documented that overconfidence in sporting ability predicted increased popularity over time. These benefits of overconfidence were extended to economic experiments in the lab. Charness et al. (2018) show that people were more likely to state higher levels of confidence when doing so would deter a competitor to enter a tournament.[2] In Schwardman and van der Weele (2019)'s experiment, half of the participants anticipate to later have to persuade peers that they performed well at a task. The authors found that participants who did expect to engage in such interactions formed positively biased beliefs about their performance at a task and that this increase in confidence led to higher expected payoffs. In the context of social status, Anderson et al. (2012) found that people are more impressed by overconfidence than actual competence and therefore overconfident individuals achieve higher social status in groups.[3]

These findings support the idea that individuals become overconfident in interpersonal interactions when overconfidence increases their expected payoffs. But what are the mechanisms at work? In Section 2, I describe a theoretical framework from evolutionary psychology which proposes that overconfidence is a form of self-deception that operates through deception.

1.2 Self-deceiving to Better Deceive Others

1.2.1 Self-deception

Self-deception is originally defined by Gur and Sackeim (1979) as a state that requires "two representations of some aspects of reality, one of them accurate and the other systematically inaccurate, and the part with access to the accurate information (the self-deceiver) must have control over the information available to the other part (the deceived self)" (Pinker, 2011, p. 36). However, there is no methodological tools from economics or psychology to show that individuals

[2]In contrast, participants were also more likely to state lower levels of confidence when being under-confident is the optimal solution (i.e., when they get higher payoffs if their competitor enters the tournament).

[3]Since this thesis focuses on strategic overconfidence, I only reviewed experiments that support the strategic aspect of holding positively biased beliefs. However, there are also evidence that people distort their beliefs in the opposite direction (Andolfatto et al., 2005). Such strategies include self-handicapping (Shepperd and Arkin, 1989; Berglas and Jones, 1978; Baumeister et al., 1989; Hoyle et al., 1999; Bénabou and Tirole, 2002; Alaoui, 2009; Ishida, 2012), strategic underconfidence (Charness et al., 2018), strategic cynism (Di Tella et al., 2015; Ging-Jehli et al., 2019) and defensive pessimism (Cantor and Norem, 1989).

hold two conflicting versions of reality. For this reason, I will use the definition by von Hippel and Trivers (2011) that relaxes this assumption and proposes that self-deception occurs if individuals strategically distort their perception of reality. As pointed out by Schwardman and van der Weele (2019), this definition *"captures both strategic and intentional shift in confidence (the "deception") and the fact that people make these beliefs their own and are willing to wager money on them (the "self")"* (p.22) and, thus, enlightens the trade-off between strategic self-enhancement and accuracy this thesis focuses on.

1.2.2 The Evolutionary Approach

Deception (i.e., the act of causing someone to believe something that is untrue) is frequently observed in empirical data (Hyman, 1989; DePaulo et al., 1996; Rosenbaum et al., 2014). Evolutionary psychologists propose that deception as emerged as a strategy in the struggle for scarce resources. Indeed, Vrij (2008) argues that one of the main reasons for people to lie is an increase in material gains. Other studies have shown that people lie to those on whom they depend to receive resources that might not otherwise be provided (Steinel and De Dreu, 2004) and 50% of people daily deceptions are intended to gain resources for one's self (DePaulo and Kashy, 1998). Standard economic theory assumes that a rational agent is honest only if the reward for honesty outweigh the one associated with dishonesty (Becker, 1968). When punishment is available, agents always lie if the benefits are high enough to cover the cost of punishment upon detection (Akerlof, 1970; Bentham, 1789; Lewicki, 1983).

However, a large body of evidence from economics shows that individuals do not always tell lies that would have increased their material payoffs.[4] Building on these evidence, some economic models now assume that lying has intrinsic costs that directly enter the utility function (Mazar et al., 2008; Fischbacher and Föllmi-Heusi, 2013; Gneezy et al., 2018). An alternative explanation that has been proposed by psychologists is that lying imposes cognitive loads on the deceiver (Vrij and Ganis, 2014; Vrij et al., 2017). When lying, the deceiver has to maintain two versions of the story in his mind (the reality and the fiction). Zuckerman et al. (1981)'s early study shows that lying can be more mentally taxing than telling the truth. fMRI research has shown that deceiving is associated with an increased activation of the prefrontal cortex (Spence et al., 2008) and execu-

[4]Examples of experimental evidence include: Abeler et al. (2014, 2016); Cohn et al. (2014); Dreber and Johannesson (2008); Erat and Gneezy (2012); Evans III et al. (2001); Fischbacher and Föllmi-Heusi (2013); Gneezy (2005); Lundquist et al. (2009); Hannan et al. (2006); López-Pérez and Spiegelman (2013); Mazar et al. (2008); Sutter (2008); Shalvi et al. (2011).

tive control processes such as working memory (Christ et al., 2008), supporting this idea that deception imposes cognitive load on the deceiver.

On the one hand, successful deception can lead to substantial benefits for the deceiver, at the expense of the deceived (DePaulo, 2004). On the other hand, individuals may let out behavioral cues of deception, leading to substantial costs (i.e. punishment) for the deceiver if the deception is detected (Boles et al., 2000; Schweitzer et al., 2006). Because of this constant co-evolutionary struggle between deceiver and deceived, self-deception may be an important deception tool that decreases the probability of the deception to be detected (Trivers, 1976, 1985, 2000, 2010). The key idea of this evolutionary approach can be found in Robert Trivers' early work:

> *"If (as Dawkins argues) deceit is fundamental in animal communication, then there must be strong selection to spot deception and this ought, in turn, select for a degree of self-deception, rendering some facts and motives unconscious so as to not betray by the subtle signs of self-knowledge the deception being practiced. Thus, the conventional view that natural selection favors nervous systems which produce ever more accurate images of the world must be a very naive view of mental evolution."* (Trivers, 1976, p. 20)

A large literature on lie-detection advocate that people are generally not good at spotting deception.[5] Bond and DePaulo (2006)'s meta-analysis reports an overall detection rate of 54% (barely above chance). However, von Hippel and Trivers (2011) and Belot and Van de Ven (2017) point out the lack of incentives to deceive or detect deception (because of the limited or nonexistent consequences of the deception on both the deceiver and the deceived), and the restricted context of the interactions in lie detection studies that may explain the low detection rate documented in the psychological literature.

Experiments with more sophisticated communication have shown to increase the rate of detection (DePaulo et al., 2003; Zuckerman et al., 1981; Belot and Van de Ven, 2017; von Hippel et al., 2016). Colwell et al. (2009) found that being trained to spot reliable cues of deception such as pupil dilatation (Wang et al., 2010) or fake smiles Ekman et al. (1988) leads to a significantly higher total accuracy (77%) than for untrained observers (57%). Mann and Vrij (2006) found an accuracy rate of 72% in police officers when asked to evaluate tapes from actual

[5]Examples of experimental evidence include: Bond and DePaulo (2006, 2008); Buller and Burgoon (1996); DePaulo et al. (1985); Ekman and O'sullivan (1991); Hartwig et al. (2004); Vrij and Mann (2005); Vrij (2008).

criminal interrogations. Diary research suggests that people detect deception at a rate that is well above chance. Participants in DePaulo et al. (1996)'s diary study reports that 15% to 23% of their lies were detected and that 16% to 23% of them were unsure that their lies were detected. Deception can also be detected by cues associated with cognitive load (Vrij, 2004; Vrij et al., 2006). The results from Vrij et al. (2017)'s meta-analysis show that imposing cognitive load on liars leads to a detection of lies of 71%. These results suggest that people's ability to detect deception might have been underestimated.

von Hippel and Trivers (2011) propose that self-deception (i) helps to hide cues of deception and (ii) deters the cognitive costs associated with lying. This idea has been formalized by Aviad Heifetz and co-authors who wrote that:

> *"In almost every game, for almost every distortion of a players actual payoffs, some extent of this distortion is beneficial to the player because of the resulting effect on opponents' play. Consequently, such distortions will not be driven out by any evolutionary process involving payoff-monotonic selection dynamics, in which agents with higher actual payoffs proliferate at the expense of less successful agents."* (Heifetz et al., 2007, p. 2)

1.2.3 The Intrapersonal Perspective

The standard psychological approach argue that overconfidence is a *"costly strategy maintained because it brings psychological benefits to overconfident individuals"* (Pinker, 2011). This argument has been supported by an extensive empirical literature in psychology showing that holding positively biased beliefs about one's self enhances self-esteem (Alicke, 1985; Dunning et al., 1995) as well as physical and mental health (Alloy and Abramson, 1979; Taylor and Brown, 1988; Taylor et al., 2000; Korn et al., 2013; Carver and Scheier, 2014). In addition, Puri and Robinson (2007) found that over-optimistic[6] individuals work more, save more, expect to retire later and are more likely to remarry after divorce.[7] I argue that these psychological advantages are by-products of the primary evolutive function of self-deception and do not conflict with Trivers (1976) argument.

[6] While overconfidence refers to positively biased beliefs of the perception of the self, overoptimism consists in overestimation of the probability of desirable events and underestimation of the probability of undesirable ones.

[7] Other advantages of overoptimism includes: lower probability of rehospitalization after surgery (Novotny et al., 2010), quicker recovery from illness (Hernandez et al., 2015) and better coping for soldiers during war time (Watson, 2006).

1.3 How do People Self-Deceive?

We have seen that people strategically bias their beliefs about themselves in social interactions when doing so leads to higher expected payoffs. But how these biased beliefs are formed and maintained? In this section, I review the various self-deceptive strategies documented in the existing literature. I first describe strategies that arise before the information is discovered (strategic ignorance). I then turn to strategies that take place after the information is discovered but not fully integrated (biased encoding processes). Finally, I report strategies that occur after the information has been encoded (reality denial).

1.3.1 Strategic Ignorance

People can prevent unwelcome information to be encoded in the brain by actively avoiding sources of information that may hold bad news (Golman et al., 2017).[8] Standard economic theory predicts that free information that can lead to better decisions should never be avoided. However, there are many situations in which people actively avoid such information, depriving themselves of potentially valuable inputs into decision-making. Individuals at risk for health conditions avoiding medical tests (sometimes at the expense of their life expectancy) is probably the most compelling illustration of active avoidance of instrumental information.[9] Out of the medical sphere, studies have shown that individuals are reluctant to obtain information concerning their beauty and intelligence (Eil and Rao, 2011; Mobius et al., 2011; Burks et al., 2010), social abilities (Trope et al., 2003), the value of their portfolios when the stock market is down (Karlsson et al., 2009; Loewenstein et al., 2016), alternative products after they made a purchase (Olson and Zanna, 1979) and even payoffs (Huck et al., 2015). Building on these empirical evidence, theoretical models by Köszegi (2006) and Weinberg (2009) assume that individuals not only consider the instrumental value of the information they receive, but also the gain or loss in utility associated with its valence (positive or negative). In this case, people are more likely to avoid self-relevant information if its content is likely to worsen the perception of their self-image.

When information is not costless to acquire (i.e. individuals have to spend time

[8] In order to be considered as self-deceptive, Golman et al. (2017) argue that information avoidance must be active in the sense that "(i) the individual is aware that the information is available, and (ii) the individual has free access to the information or would avoid the information even if access were free."

[9] Examples of information avoidance regarding health include: Lyter et al. (1987); Lerman et al. (1996, 1999, 2002); Sullivan et al. (2004); Yaniv et al. (2004); Dawson et al. (2006); Thornton (2008); Sweeny et al. (2010); Oster et al. (2013); Ganguly and Tasoff (2016).

on searching information or pay a fee to obtain it), people can favor sources of information that are consistent with their own view (Frey, 1986). In line with this idea, studies have shown that people stop gathering information when they like the early returns but keep gathering when they do not (Ditto and Lopez, 1992). People may also choose to allocate their attention differently between positive and unfavorable or threatening information. For example, people have been shown to spend more time looking at negative rather than positive information about their potential partner when they expect to be rejected (Wilson et al., 2004) and more eager to remove a background noise on a speech that was consistent with their view than when it was not (Brock and Balloun, 1967). Eye-tracking studies also reveal that people pay more attention to aspects of the available information that they would prefer to be true (Isaacowitz, 2006; Isaacowitz et al., 2008).

1.3.2 Biased Encoding Processes

When unwanted information is nevertheless encoded, people can still filter out negative information (Taylor and Brown, 1988) and favor evidence that they are motivated to believe (Lord et al., 1979; Dawson et al., 2006; Babcock et al., 1995b). A growing body of empirical evidence support the idea that people tend to put more weight on good news than bad news (Eil and Rao, 2011; Mobius et al., 2011; Sharot et al., 2012; Wiswall and Zafar, 2015; Sharot and Garrett, 2016).[10] In Mobius et al. (2011) and Eil and Rao (2011) experiments, the authors found that subjects react more to positive feedback about their relative performance and do not update sufficiently negative ones. Others have shown that people give more weight to information that support their own view.[11] When presented with arguments that support both their view and the opposite view of a discussion, people with strong priors about one side or the other tend to end up the discussion even more polarized than when they started (Lord et al., 1979; Dawson et al., 2002; Glaeser and Sunstein, 2013; Sunstein et al., 2016).

1.3.3 Reality Denial

Even when people attend to unwelcome information and encode information in an unbiased manner, such information can be strategically forgotten or misremembered (Crary, 1966) because it may help think of past behavior under a

[10] On the contrary, Kuhnen (2015) found that investors tend to react more to information infer from losses than gains.

[11] Examples of confirmation bias include: Wason (1968); Rabin and Schrag (1999); Jones and Sugden (2001); Descamps et al. (2016); Charness and Dave (2017).

positive light (Moore, 2016). Theoretical models show that individuals have an incentive to forget undesirable signals (Bénabou and Tirole, 2002) and recall negative signals with probability below the actual percentage (Gottlieb, 2014). These predictions have been supported by numerous experimental results. Early psychological findings showed that people bias their memory of their beliefs about their own skills/attributes when it comforts them in the idea that they have improved (Conway and Ross, 1984; Croyle et al., 2006). People are also more likely to remember their success rather than their failure (Korner, 1950; Mischel et al., 1976), good things rather than bad things (Thompson and Loewenstein, 1992; Story, 1998; Sedikides and Green, 2009) and when they behaved ethically rather than when they did not (Kouchaki and Gino, 2016).

Evidence from economics tend to show that individuals manipulate their memories to preserve their self-image. Individuals forget more about their incorrect answers than correct ones of past performance in an IQ test (Li, 2017; Chew et al., 2018; Zimmermann et al., 2019). In the social domain, Shu and Gino (2012) found that cheaters recalled fewer items from a moral code than non-cheaters. Li (2013) found that betrayed trustors in a trust game recall less accurately the outcome of the game compared to trustors who benefited from an altruistic decision from their trustee. Saucet and Villeval (2018) found that dictators in a dictator game were more likely to remember their altruistic than their selfish decisions.

Finally, if information is accurately recalled, people can use *self-signaling* strategies (Bénabou and Tirole, 2016) by which they "produce" signals that allows them to later interpret their prior choices as desirable (Quattrone and Tversky, 1984; Bodner and Prelec, 2003; Bénabou and Tirole, 2004, 2011) or rationalize the motives behind the behavior. To illustrate the latter, Von Hippel et al. (2005) showed that when cheating can be seen as unintentional, people who showed a self-serving bias in another domain were more likely to cheat. People are also more likely to avoid sitting next to disabled people (Snyder et al., 1979) or helping African Americans (Saucier et al., 2005) when their behavior can be rationalized.

1.4 Aim and Outline

A growing body of evidence now support Trivers (1976)'s idea that individuals strategically self-deceive in social interactions because doing so leads to superior payoffs. However, evidence of actual self-deception and its causal effect are scarce in economics and our understanding of this phenomenon remains limited. First, few experimental designs allow to measure to what extent people believe what they claim to believe. Most results from the psychological literature rely on self-

assessment and do not provide direct monetary incentives for accurate reporting (or cost for not being accurate). The same concern arises when those claims are made public.

Second, despite these evidence, overconfidence has also been shown to have undesirable economic consequences. Overconfident workers choose risky payment scheme that earn them less than a lower fixed piece-rate (Barron and Gravert, 2018), overconfident male traders take inconsiderate risks (Barber and Odean, 2001) and homeowners have unrealistic expectations regarding the value of their property in the future (Case and Shiller, 2003). What is even more problematic is that the costs associated with decisions made by overconfident individuals can also affect others. As Johnson and Fowler (2011) remind us:

> *"Overconfidence has been blamed throughout history for high-profile disasters such as the First World War, the Vietnam war, the war in Iraq, the 2008 financial crisis, and ill-preparedness for environmental phenomena such as Hurricane Katrina and climate change."*

In finance, Heaton (2002) and Malmendier and Tate (2005) found that overconfident CEOs make poor investments or mergers decisions. Overconfidence also leads to excessive entry into markets (Camerer and Lovallo, 1999) and inefficient delays in bargaining settings (Ortner, 2013). In psychology, Shipman and Mumford (2011) found that overconfident leaders are prompt to pursue personalized objectives at the expense of others, make mistakes and fail to identify deficiencies and problems. In the shade of these costs, understanding why overconfidence is beneficial in some situations but detrimental in others appears fundamental.

Across the three essays constituting this thesis, I aim to address this question by proposing a series of experiments in which in which (i) participants' beliefs are exogenously manipulated; (ii) the discrepancy between participants' beliefs and participants' actual performance is clearly measured; and (iii) holding self-serving biased beliefs has a direct cost. These features allow to causally identify the individual and social impact of overconfidence.

Chapter 2 contributes to the literature by (i) providing compelling evidence that overconfidence emerges in social interactions primarily when it provides an advantage; and (ii) showing that the degree to which people can self-deceive depends on the mental wiggle room they have to do so. In this experiment, participants are incentivized either to form accurate beliefs about their performance at a test, or to convince a group of other participants that they performed well. I also vary participants' ability to freely gather information about their performance. I

show that participants who expect to have to convince others about their ability selectively search for information in a way that is conducive to receiving more positive information on their performance. Results also show that participants who expect to have to convince others are more overconfident than those who do not; and this increase in confidence has a positive effect on their persuasiveness.

The second part of this thesis focuses on two types of interactions in which individual self-interest conflicts with social interest: bilateral negotiations (Chapter 3) and the selection of leaders (Chapter 4). Chapters 3 and 4 contribute to the literature by showing that in such interactions, overconfidence is likely to be socially detrimental, while the benefits of being overconfident for the overconfident individual are sustained.

In Chapter 3, I investigate the effect of relative confidence on the outcome of a bilateral bargaining situation. Participants are matched in pairs and each pair performs a task to earn points that are allocated to a group account. At the end of the task, I manipulate participants' confidence about their performance relative to their partner's. The group account is then divided in two unequal shares (70%-30% percent of the group account) and each pair is asked to agree on how to allocate the shares among its members. I use a 3-stage negotiation process in which pairs of participants are given an opportunity to reach an agreement in each of those 3 stages. However, the longer each pair takes to agree, the lower the final amount of each share will be. If the pair fails to reach an agreement, both members end up empty-handed. I found that an increase in confidence at the pair level lowers the social outcome of the negotiation by increasing the likelihood of impasses and delays during the negotiation process. In contrast, results shows that negotiators who are relatively more confident than their partners are more likely to end up with larger payoffs.

In Chapter 4, I investigates whether individuals become strategically overconfident when being a leader give access to valuable resources and to what extent decisions made by overconfident leaders are detrimental to the group. In this experiment, participants undertake an effort task and are then matched in groups of four. Each group have to select a leader that will make a series of risky binary choices on behalf of the group. For each decision, the probability to obtain the desirable outcome depends on the likelihood that the leader was ranked as the best performer in his group. Thus, selecting the best performer of the group as the leader maximizes the social outcome. Before making their choice of leader, group members are allowed to communicate among them via a group chat. Half of the leaders are offered a bonus to be the leader. Results show that leaders in

the treatment with the bonus (i) are less likely to be the best performer in their group and (ii) make overconfident choices that earn less money for their group compared to leaders who did not receive the bonus.

These findings enhance our understanding of the situational determinants of overconfidence and bring an explanation to why overconfidence persists, even when it leads to dramatic consequences. Even though overconfidence can be socially costly, there is no reason to expect individuals to be well-calibrated when being overconfident remains beneficial for the overconfident individual. Most real-life interactions involve situations where agents private interest is orthogonal to the interests of other agents engage in the interaction: markets, tournaments, public goods, etc. Hence, these insights should help organizations to anticipate situations where overconfidence is likely to appear and, more importantly, when it is likely to be socially costly. Chapter 5 discusses the scope of these results, as well as the possible extensions of this thesis, and concludes.

Chapter 2

Strategically Delusional[1]

2.1 Introduction

For a long time, most standard theories in economics (Von Neumann and Morgenstern, 1953) and psychology (Maslow, 1950; Körding and Wolpert, 2004) have assumed that people collect and process information in a way that gives them an accurate perception of reality. But empirical research has shown that most people are actually overconfident regarding their own abilities. They believe that they are more skilled, more attractive, and in better health than others (Svenson, 1981; Gabriel et al., 1994; Weinstein, 1980; Epley and Whitchurch, 2008). This overconfidence occupies a particular place in the collection of behavioural biases. The widespread presence of inflated self-beliefs presents an instance where people are not just making random mistakes, it appears instead as a systematic tendency to venture in self-serving delusions.

Many psychological studies have suggested that overconfidence arises because it has a consumption value (Taylor and Brown, 1988): people enjoy basking in the belief that they are better than they actually are. However, miscalibrated beliefs have a cost. The perception of our own abilities/attributes influences how we make decisions. Thus, the outcome of the decisions we make depends - in part at least - on being able to accurately evaluate our own abilities (Dunning et al., 2004). If overconfidence can lead to costly mistakes, an adequate explanation of its prevalence and persistence most likely requires for it to provide some material benefits as well. In the present paper, we investigate the idea that overconfidence emerges as a strategy to gain advantages in social interactions.

[1] Co-authored with Changxia Ke, Lionel Page and William von Hippel. Accepted at *Experimental Economics*.

To address this question, we design an experiment to investigate (i) whether overconfidence is more likely to emerge in situations where people anticipate the need to convince others about their performance and (ii) whether this strategically motivated overconfidence helps them be more persuasive. In service of these goals, we use a 2x3 design in which we manipulate participants' anticipation of strategic interactions and their opportunity to gather information about their performance. Participants were first asked to complete a general-knowledge test. At the end of the test, half of the participants were initially incentivized to give an accurate estimate of their performance ("Accuracy Task"), while the other half was initially incentivized to convince other participants that they performed well on the test ("Persuasion Task"). Participants who completed the Accuracy Task were next instructed to complete the Persuasion Task, and vice versa. Participants were not given any information about the second task until they finished the first one.

This design allows us to compare beliefs of participants who initially attempted to persuade others of their strong performance to the beliefs of those who initially attempted to assess their performance accurately. We cast a light on how these beliefs are formed by examining how participants engage in information sampling when they have the freedom to self-select information, compared to when they are either given no information about their performance or when the information is selected by the experimenter. The goals of this design are to study whether overconfidence emerges in anticipation of the need to persuade, whether the need to persuade leads to biased information gathering, whether biased information gathering facilitates the emergence of overconfidence, and whether overconfidence in turn facilitates persuasion.

We first conducted this experiment on the online platform of Amazon Mechanical Turk (in Study 1) and then replicated a subset of conditions in the controlled laboratory environment (in Study 2) as a robustness check for our findings. Across the two studies, we find that participants show greater confidence in their performance when they initially anticipate the need to persuade others rather than to appraise themselves accurately, although their performances are very similar. Furthermore, participants who initially intend to persuade engage in biased information sampling in a manner that facilitates self-confidence. We also find evidence from the laboratory study suggesting that holding positively biased beliefs may help individuals in their effort to be more persuasive. Overall, these results provide support for the idea that overconfidence arises in social interactions when it can provide a strategic advantage.

Our paper broadly relates to research on motivated beliefs (Bénabou and Tirole, 2016). The idea that beliefs can be used strategically to achieve higher payoffs has been formalized in economics by Bénabou and Tirole (2002). In particular, Hong et al. (2018) model motivated beliefs as a way to alleviate the underinvestment problem associated with present bias, which has been supported by some empirical evidence (Puri and Robinson, 2007; Vialle et al., 2011). These first economic models of motivated beliefs focus on the *intrapersonal* advantages of holding positively biased beliefs. Another strand of the literature proposed instead that motivated beliefs can provide an *interpersonal* advantage. Agents who are overconfident may be more effective at signalling their ability in strategic situations where agents' true types cannot be observed (von Hippel and Trivers, 2011; Bénabou and Tirole, 2016). Our paper contributes to this literature and presents evidence that overconfidence provides interpersonal benefits because it advantageously influence others in social interactions.

Building on Trivers (1976)'s hypothesis that self-deception evolved to deceive others more effectively, von Hippel and Trivers (2011) propose that overconfidence plays an interpersonal role by enhancing others' perception of one's positive qualities. While bluffing may be sufficient to deceive others, self-deception provides additional benefits. First, self-deception may alleviate the cognitive costs of deception (e.g. holding in your mind two competing versions of the reality, the one you believe in and the one you want to impart to others). Second, if the cognitive costs of deception generate visible cues of deception (e.g. being slower when generating arguments), self-deception could be a way to avoid such cues.[2] The idea that an evolutionary process can lead agents to generate mis-calibrated beliefs when one agent's beliefs can influence others in social interactions has been formalised by Heifetz et al. (2007). The authors show that there is no reason to expect an evolutionary process to lead to agents having well-calibrated beliefs, as soon as agents' beliefs can influence others' perceptions in social interactions.

Our paper fits in the emerging strand of empirical research investigating the interpersonal advantages of overconfidence. In a series of experiments, Anderson et al. (2012) provide evidence that individuals can attain a higher status in social interactions when they are overconfident. A similar effect is found by Murphy et al. (2017), who tracked 894 high school boys across two school years, and documented that overconfidence in sporting ability predicted increased popularity over time. These social benefits of overconfidence were extended to an economic experiment in the lab, in which Charness et al. (2018) found that participants

[2]Self-deception itself can incur other costs, like most other strategies. The decision whether to engage in it depends on how benefits weigh against costs.

were more likely to overstate claims of their strengths when it is optimal to deter opponents' entry in a contest. They also find evidence that these overstatements may not be fully self-aware.

The closest study to ours in the literature is Schwardman and van der Weele (2019). In their experiment, the authors elicited participants' beliefs about their performance after an intelligence test. The beliefs were elicited twice (once right after the test and before participants were given a noisy signal about their performance, and once after observing the signal). Prior to estimating their performance, half of the participants (in the treatment group) were informed that they would undertake a face-to-face interview afterwards to persuade interviewers that their performance belonged to the top 50% of their group, whereas the other half (in the control group) were not given this information. The authors find two results in support of the strategic-overconfidence hypothesis. First, those who expected to convince others of their superior performance were more overconfident than those who did not have this expectation. Second, they find that their overconfidence appeared to make them more persuasive in the face-to-face interviews.

Our design differs from Schwardman and van der Weele (2019) primarily in that, instead of providing a noisy signal about their performance to generate exogenous variations in confidence, we elicited participants' beliefs under different feedback conditions to investigate the underlying mechanism whereby people deceive themselves (i.e. how people form and maintained biased beliefs about themselves).[3] This feature of our design is similar to that of Smith et al. (2017) who investigated whether people gathered more or less information as a function of the congruence of the information with their persuasive goals. Participants in our research were given the opportunity to choose which information they wanted, rather than the decision of whether to discontinue their information search early (i.e., information avoidance). We find that people bias their information search with regard to content, in addition to the findings of Smith et al. (2017) regarding the amount of information. Our study also investigates if people engage in this biased search in a manner that helps them strategically bolster their confidence, whereas Smith et al. (2017) did not examine overconfidence.

[3]There are also other differences in design details. For example, in the treatment group, Schwardman and van der Weele (2019) elicited participants' beliefs after they had been informed about whether they were going to undertake an interview. Participants who undertake the interview face a trade-off between being accurate now and being convincing in the future while reporting their beliefs. Our design tries to avoid such trade-offs by having half of the participants anticipate a need to be accurate and the other half anticipate a need to be persuasive. Instead of having a face-to-face interview, our participants then wrote an essay to convince the reviewers without any direct interactions.

Corroborating the findings in Schwardman and van der Weele (2019), we provide complementary evidence on the benefit of strategic overconfidence. When our participants were motivated to persuade others about their superior performance they tended to be more confident than when their goal was to report the most accurate beliefs. This additional confidence also appeared to make them more convincing, as they received higher assessments from the reviewers. We also find that participants who are motivated to persuade engage in biased information search in a manner that is conducive to receiving more positive information about their performance. Altogether, the results of both studies suggest that agents are not as naively delusional as has been suggested. Instead, they may often be "strategically delusional", forming unrealistic beliefs that give them an advantage in social interactions.

The remaining sections are organised as follows. Section 2 details the experimental design and hypotheses. In Section 3, we describe the procedures and display our main results. Section 4 presents our conclusions.

2.2 Experimental Design and Hypotheses

2.2.1 General Design

We design an experiment to investigate whether the propensity to engage in self-deception is sensitive to the potential advantage of holding overconfident beliefs in a given strategic interaction. We first present participants with a timed general-knowledge test in the form of a multiple-choice questionnaire.[4] The test is composed of 30 questions of moderate difficulty, as indicated by the percentage of participants who answered the questions correctly in a previous MTurk experiment (Murphy et al., 2015).[5] Participants have 15 seconds to answer each question and do not receive any information regarding the following parts of the experiment at this point.

We then vary participants' anticipation of strategic interactions by asking participants to undertake two incentivized tasks sequentially: an Accuracy Task and a Persuasion Task. In the Accuracy Task, participants are incentivized to give their best guess about their absolute and relative performance. In the Persuasion Task, participants are incentivized to convince others that they performed

[4] We provide full experimental instructions in Appendix A.4.

[5] A 31st question served as an attention check, and was designed such that all participants should be able to answer it correctly if they paid attention to the question.

well in the general-knowledge test. We design the experiment such that half of the participants join the "*Accuracy-first*" treatment (i.e., are presented with the Accuracy Task first, which is then followed by the Persuasion Task), while the other half of the participants join the "*Persuasion-first*" treatment, and do both tasks in the reverse order. Participants are only informed about the nature of each task as they undertake it.[6] This manipulation of task order enables us to observe participants engaging in the same set of tasks, but varies whether they are initially incentivized to form accurate beliefs about their performance or to form beliefs in a manner that would help them to convince others of their strong performance. This design enables us to investigate how participants form their beliefs and how their ability to persuade others is affected by the anticipation of different types of interactions.

Accuracy Task: Participants undertake two independent belief elicitation tasks. First, they are asked to give an estimate of the number of correct answers they achieved in total (i.e., absolute performance). They are then asked to give an estimate of how well they did compared to other participants in the study (i.e., relative performance).[7] For the first belief elicitation, we asked participants to guess how many questions they believe they answered correctly in the test on a scale from 0 to 31 (thereby including the final, attention-check question, which was not presented separately from the rest of the questions). For the second belief elicitation, participants were asked to give an estimate of the percentage of participants whom they believe they outperformed on a scale from 0 to 100%. To avoid potential hedging behaviour between the two belief elicitations, the questions were presented sequentially.[8]

Persuasion Task: Participants are asked to convince a group of reviewers about the strength of their performance by writing a short essay.[9] They are told that the reviewers are another group of participants who did not take the test. Participants are informed that reviewers will be reading their essays and will rate

[6]Hence, in the *Persuasion-first* condition, participants did not know that following the Persuasion Task they would be rewarded for accuracy. As such, they were not in a position to engage in a cost-benefit analysis regarding whether they should self-deceive as there is only perceived benefit of being overconfident if they believe this will help them in persuading the reviewers. Similarly, in the *Accuracy-first* treatment, there is no incentive for participants to over-report their performance as they were not yet aware of the upcoming Persuasion Task.

[7]Note that the reference group for comparison is everyone else in Study 1 (which is 583 MTurk participants), and this was made clear in our instructions. Of course, people would not have known exactly how many other participants there would be, and hence probably drew up an image of the average MTurk workers in their mind.

[8]The high correlation between both measures in our data suggests that hedging behaviour was unlikely in our experiment (Pearson correlations: r_s=0.73 in Study 1 and r_s=0.71 in Study 2; p <0.001 for both studies).

[9]Examples and analyses of these essays are provided in Appendix A.3.

them on: (i) how many questions they believe the participant answered correctly and (ii) how convincing they think the essay is. Each essay is reviewed by five different reviewers independently. By rewarding participants in part based on the reviewers' estimate of their performance, we attempted to ensure that participants do not have an incentive to be convincing by claiming they performed badly at the task.[10]

Incentives scheme: The presence of different incentives schemes between the Accuracy Task and the Persuasion Task could affect how much attention participants give to the feedback and their selection of information in the *self-chosen information* condition. For this reason, we use a reward scheme based on relative performance within each reference group in both the Accuracy Task and the Persuasion Task to ensure an identical payoff structure across treatments.[11] For both elicitations in the Accuracy Task, participants are rewarded if they give more accurate estimates than other participants. If participants' estimates are among the top 10% of the most accurate estimates, they receive \$2; If participants' estimates are among the top 50% (but below the top 10% of the most accurate estimates), they receive \$1.
In the Persuasion Task, participants are rewarded based on comparisons of the average ratings given by all five reviewers. Participants were told that if the reviewers' average assessment of their own absolute performance is in the top 10% of the (average) ratings within their comparison group, they receive \$2; If the reviewers' assessment was in the top 50% (but below top 10%), they receive \$1. Similarly, if participants' essays are rated in the top 10% of the most convincing essays within the comparison group, they receive \$2; if participants' essays are rated in the top 50% (but below top 10%) of the most convincing essays, they receive \$1.[12]

Because data collection was continuous and could not be broken down to sessions, participants were compared against all the MTurk workers who participated in the study for payment. Participants were told at the end of the experiment that they will receive their payment for the Accuracy Task within three days and their payment for the Persuasion Task within two weeks.[13]

[10] Section A.3.3 in the Appendix provides further evidence that most participants did not engage in such behaviors.

[11] While this mechanism might engage participants' second order beliefs, the monotonicity of the reward scheme relative to accuracy does not change across treatments (i.e., regardless of their second order beliefs, it pays more to be more accurate).

[12] Reviewers were paid a fixed wage to rate one essay only. Participants were not given information about how the reviewers' payment would be calculated.

[13] Respectively, those were the time frames anticipated to complete the data collection from the main participants and the reviewers.

Information Feedback: To further investigate how participants form beliefs about their performance, before proceeding to the Accuracy or Persuasion Task, we also vary the information available to participants about their performance. We use three different information conditions: a *No Information* condition, a *Given Information* condition and a *Self-Chosen Information* condition.[14] In the *No Information* condition, participants do not receive any external feedback about their performance after the test. In the *Given Information* condition, participants are shown 10 pre-selected questions and whether they answered them correctly or not.[15] The 10 questions are selected to reflect the general level of difficulty of the 30 proper general knowledge questions.[16] These questions are the same for all participants. By virtue of our question sampling, the percentage of correct answers shown to participants in this feedback should predict the overall percentage of correct answers they are likely to receive for the entire test. In the *Self-Chosen Information* condition, participants are presented with the list of all the questions they faced during the test (excluding the last item that was used as an attention check). The questions appear in random order and participants are told they are to select 10 questions of their choice to check whether they answered them correctly. As in the *Given Information* condition, participants in the *Self-Chosen Information* condition are informed for each selected question whether their answer was correct. The only difference between the two conditions lies in whether the questions were selected by the participants or the experimenter.

Factorial design: We cross the information conditions and the treatments in a 2x3 design represented in Table 4.2. Figure 2.1 provides the structure of this design in a flow chart.

Table 2.1: A 2X3 experimental design.

	Information conditions		
Treatments	*No Information*	*Given Information*	*Self-Chosen Information*
Accuracy-first	*NI* x *Acc.1st*	*GI* x *Acc.1st*	*SCI* x *Acc.1st*
Persuasion-first	*NI* x *Per.1st*	*GI* x *Per.1st*	SCI x *Per.1st*

Notes: Table 4.2 displays the six cells of our 2X3 factorial design. NI stands for the *No Information*, GI for the *Given Information* condition, SCI for the *Self-Chosen Information* condition. *Acc.1st* refers to the Accuracy-first treatment and *Pers.1st* refers to the Persuasion-first treatment.

[14] For the sake of clarity, we refer to *Accuracy-first* (*Acc.1st*) and *Persuasion-first* (*Per.1st*) as "treatments" and *No Information* (*NI*), *Given Information* (*GI*), and *Self-Chosen Information* (*SCI*) as information "conditions".

[15] Participants are not told what the correct answers are if their answers are wrong.

[16] These 10 questions were chosen according to the accuracy rate of each question in an experiment run by Murphy et al. (2017) using the same knowledge test and a sample from the same population.

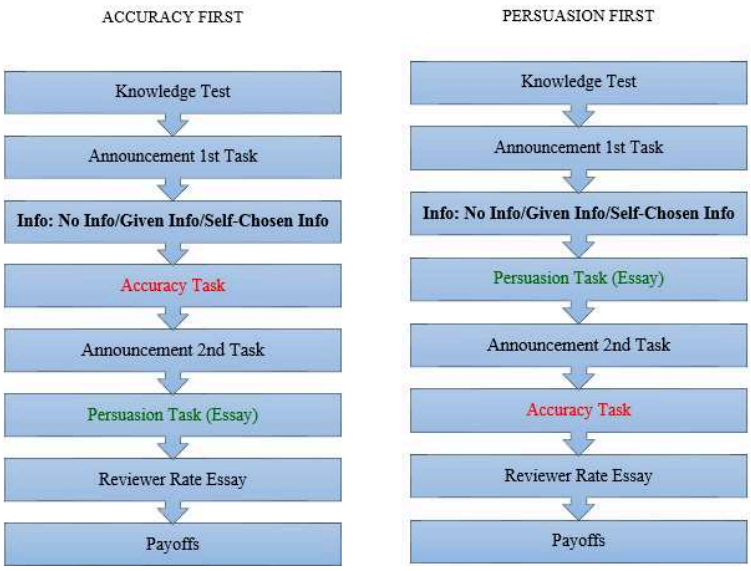


Figure 2.1: A timeline of the two different treatments.

At the end of the experiment, we recorded a range of individual characteristics to use as control variables in our analyses: participants' sex and age, as well as their dispositional overconfidence using the Over-Claiming Questionnaire (OCQ).[17]

2.2.2 Hypotheses

If overconfidence provides a strategic advantage in social interactions, we may expect that its emergence is influenced by the existence of possible gains from being more confident. The two treatments we designed induce different incentives for participants to engage in motivated reasoning. In the *Accuracy-first* treatment, participants are initially incentivized to be accurate while in the *Persuasion-first* treatment participants are initially incentivized to be persuasive. If overconfidence facilitates persuasion about one's positive qualities/attributes, there are gains from forming more confident beliefs in the *Persuasion-first* treatment. Schwardman and van der Weele (2019) already suggests that this hypothesis might be true. Hence, we expect to observe that participants in the *Persuasion-first* treatments are more likely to be overconfident than participants in the Accuracy-first treatment about their own performance and their relative position in the group. Importantly, if their overconfidence is self-deceptive rather than just bluffing, it should be carried forward to their judgments made on the following Accuracy Task, even though at that point accuracy is incentivized. In contrast, people in the *Accuracy-first* treatments should show minimal beliefs

[17]Dispositional overconfidence is the idiosyncratic trait level of overconfidence, as opposed to situational overconfidence (here, the knowledge test). We use the 25-item versions of the test proposed by Bing and Davidson (2012). See Appendix A.5 for further details.

distortion, due to the benefits of evaluating their performance dispassionately in the initial Accuracy Task. This leads to our first hypothesis.

Hypothesis 1 (Strategic confidence) *Participants in the* Persuasion-first *treatment will form more confident beliefs about their absolute and relative performance than participants in the* Accuracy-first *treatment.*

We also conjecture that it is easier for participants to distort their perception of their own performance when they have greater freedom in gathering information to form their beliefs. This hypothesis follows the insights from the literature on mental wiggle room in games where self-signalling can play a role (Grossman and Van Der Weele, 2017). In the *Self-Chosen Information* condition, participants are given the most freedom to engage in motivated reasoning and form the most favourable views about their performance. We expect participants to gather information in a biased way when given the opportunity to choose freely and when there are strategic incentives for being more confident. Specifically, we expect participants in the *Persuasion-first* treatment to select more questions they believe they have answered correctly compared to participants in the *Accuracy-first* treatment, because doing so will give them more positive feedback, which will help them form more confident beliefs and be more persuasive.

It is ex-ante unclear under which condition the wiggle room is bigger between the *No Information* and *Given Information* condition. The *Given Information* condition provides participants with information that should be representative of their overall performance during the test. However, the mechanism most often assumed to generate overconfidence is the "selective updating" of beliefs, whereby good news is weighted more heavily than bad news when revising beliefs (Eil and Rao, 2011; Mobius et al., 2014; Sharot et al., 2011; Kuhnen, 2015; Wiswall and Zafar, 2015; Bénabou and Tirole, 2016). From that perspective, receiving some information (in the *Given Information* condition) may be more conducive to self-deception because it allows participants to select and interpret the evidence in a manner that is conducive to asymmetric updating of their beliefs. In contrast, the dearth of information in the no-information condition may make this process harder. We therefore conjecture that the *Given Information* condition should give slightly more freedom to engage in motivated reasoning than the no-information condition.

In summary, we hypothesize that participants will become more confident when they are incentivized to persuade than when they are incentivized to be accurate

and the discrepancy between the two treatments should increase when participants have the opportunity to shape the feedback they receive from the test. This logic leads to the following two hypotheses.

Hypothesis 2 (Selective/biased information search) *Participants in the* Self-Chosen Information *condition will engage in selective/biased information search in a manner that is conducive to forming more confident beliefs (i.e., by sampling easier questions) when in the* Persuasion-first *treatment than when in the* Accuracy-first *treatment.*

Hypothesis 3 (Mental wiggle room & strategic confidence) *The difference between the* Persuasion-first *treatment and the* Accuracy-first *treatment (in beliefs about absolute and relative performances) will increase from the* No Information *condition to the* Given Information *condition, and increase further in the* Self-Chosen Information *condition.*

Furthermore, holding more confident beliefs will provide an advantage to participants in their effort to convince reviewers that they did well on the test. We thus propose the following hypothesis.

Hypothesis 4 (Effectiveness of strategic confidence) *More confident beliefs generated through motivated reasoning will help participants be more successful at persuading reviewers to rate them favourably.*

We pre-registered this design and hypotheses on the Open Science Framework.[18] Figure 2.2 below summarises how our main hypotheses fall within von Hippel and Trivers (2011)'s theory.

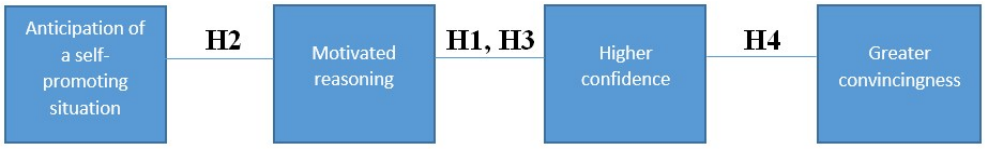


Figure 2.2: A summary of von Hippel and Trivers (2011)'s theory.

[18]The hypotheses' statements were improved for exposition purposes. The pre-registration can be found at the following link: https://osf.io/z5266/?view_only=e26aeeef9d794b9c8a91887d57323c53

2.3 Data and Results

2.3.1 Study 1: MTurk Experiment

We first implemented our 2x3 factorial design online via Amazon MTurk, where 600 individuals participated in the main part of the experiment (100 in each treatment) and 3000 others participated as reviewers.[19] The main participants were randomly allocated to one of the six treatment-conditions. Most of the participants finished the tasks within 35 minutes, and on average they earned 3.25 USD (s.e. = 0.84) plus a fixed payment of 2 USD. Reviewers in this study only received one essay each and were paid a fixed amount of 0.25 USD for an average of 5 minutes spent reviewing the essay.[20] Because our experiment is based on a knowledge test validated on US participants, only native English speakers from the USA were invited to join our study. The experiment was programmed using Qualtrics.

Results

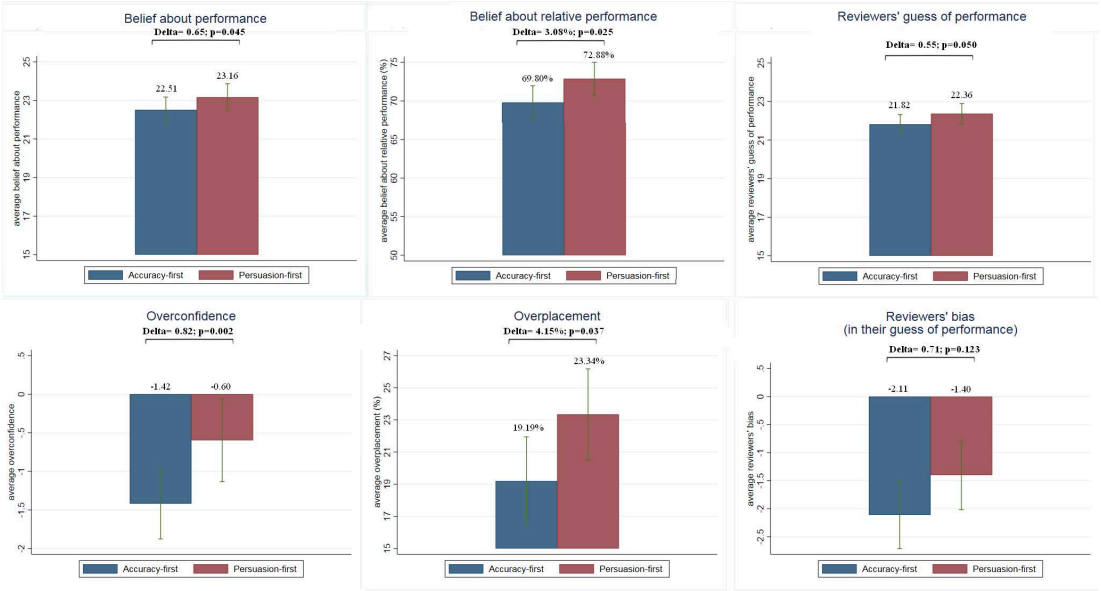


Figure 2.3: A summary of our main variables of interest comparing *Accuracy-first* treatment (left bars) and *Persuasion-first* treatment (right bars) with mean values, treatment differences and p-values (from two-sided Mann-Whitney rank-sum tests) indicated on top. All information conditions in Study 1 are pooled together.

[19] 16 observations were excluded from the analyses based on the following criteria: participants (i) provided empty essays or (ii) failed both control questions. The analyses presented in the subsequent sections are therefore conducted on the remaining 584 participants only.

[20] Participants were allowed to keep the experiment window open on their internet browser for a few hours without disconnection, so several participants took much longer.

We first hypothesise that participants in the *Persuasion-first* treatments hold more confident beliefs about their performance than in the *Accuracy-first* treatments. While we predict that this effect will be moderated by our information conditions, the direction of the effect is expected to be the same for all conditions. Hence, we first examine the discrepancy in participants' beliefs and reviewers' assessment of their performance between the *Persuasion-first* and the *Accuracy-first* treatment across all information conditions. Figure 2.3 displays the summary statistics of these measures when all information conditions are pooled together. Each bar represents the mean of each main variable of interest (with confidence intervals). The top panels display participants beliefs about their absolute and relative performance, and the reviewers' average guess of their performance. The bottom panels show the corresponding "biases" in these beliefs. We use the difference between participants' beliefs about their absolute performance and their actual performance as a measure of "*Overconfidence*", and the difference between participants' beliefs about their relative performance (i.e, the percentage of people they have outperformed) and their actual relative performances to measure "*Overplacement*".[21] We also use the difference between reviewers' average guesses of participants' performance and their actual performance to measure the biases in reviewers' estimates. The blue/left bar (in each pair-wise comparison) represents the value for the *Accuracy-first* treatment and the red/right bar represents the value for the *Persuasion-first* treatment. On top of the bars, mean values, the mean treatment differences, and the p-values from two-sided Mann-Whitney rank-sum tests are also provided in Figure 2.3.[22]

Figure 2.3 shows the following regularities on participants' and reviewers' beliefs and biases in these beliefs. First, we find participants in the *Persuasion-first* treatment overall hold more confident beliefs about their performances, compared to the *Accuracy-first* treatment - their beliefs are around 0.65 higher (p=0.045), even though their actual performances are very similar across treatments (23.93 *vs.* 23.77, p=0.763). If we further examine biases in beliefs, we find that overall the "*Overconfidence*" measure is 0.82 higher (p=0.002) in the *Persuasion-first* treatment, compared to the *Accuracy-first* treatment. Second, beliefs about their relative performances show similar patterns. Participants in the *Persuasion-first* treatment believe that they have outperformed 72.88% of the other participants, while those in the *Accuracy-first* treatment believe they have outperfomed 69.80% of the people. The treatment difference (3.08%) is significant at 5% (p=0.025).

[21] We compute participants' actual percentile by ranking them according to their performance at the test. We use a cumulative distribution function to randomly break tied performances and allocate each participant to a unique percentile.

[22] A full summary table with mean values and standard errors for all the variables is also provided in Table A.1 in the Appendix.

When the "*Overplacement*" measure is assessed, the average value is 4.15 percentage points higher (23.43% vs. 19.19%, p=0.037).[23] Finally, reviewers' guess of performances are 0.55 higher in the *Persuasion-first* than in the *Accuracy-first* treatment (22.36 vs. 21.82, p=0.005), however, when we look at the difference between reviewers' guess of performances and participants' actual performance, the differential bias in reviewers' guess is no longer significant (0.71, p=0.123). These results together support Hypothesis 1 that participants will show strategic overconfidence when motivated to persuade, but they do not address the possible causal role of participants' overconfidence on persuasiveness.

Result 1 (Strategic confidence) *Participants in the Persuasion-first treatment form more favourable beliefs about their absolute and relative performance than participants in the Accuracy-first treatment, even though their actual performance is similar.*

If the expectation of having to convince others leads to strategic self-deception, we would expect it to be reflected not only in participants' final beliefs but also indirectly in how participants choose to process information in order to form favourable beliefs. There are two conditions where participants observe information about their performance in our experiment, the *Given Information* condition where participants do not choose what information they receive and the the *Self-Chosen Information* condition in which participants choose the questions for which they want to receive feedback. We expect that participants in the *Persuasion-first* treatment will selectively choose questions they are more likely to have answered correctly (compared to those in the *Accuracy-first* treatment) in order to facilitate positive feedback. This approach may allow them to sustain a more positive belief about their performance. Our measure of "Feedback" (i.e., the proportion of correct answers contained in the 10 pre-selected questions) presented in Figure 2.4 (left panel in the last row) is consistent with this prediction.

We observe that participants in the *Self-Chosen Information* condition chose a set of questions with on average 12% more correct answers (79.69% in the *Persuasion-first* treatment *vs.* 67.4% in the *Accuracy-first* treatment, $p < 0.001$).[24] In contrast, by virtue of the experimental design, there should be no difference in this measure in the *Given Information* condition between the *Persuasion-first* and

[23] Overall, our participants slightly underestimate their absolute performance but substantially overestimate their relative performance. These results are consistent with previous studies that find underconfidence and overplacement emerges jointly depending on task difficulty (Moore and Healy, 2008; Larrick and Soll, 2007). Hence, we will only focus on the treatment differences in the following analysis.

[24] This behavior indicates a violation of Bayesian thinking. However, our design does not allow to identify the precise mechanism.

Accuracy-first treatments, as only random variation would cause the feedback to vary on the same 10 questions across the two treatments.

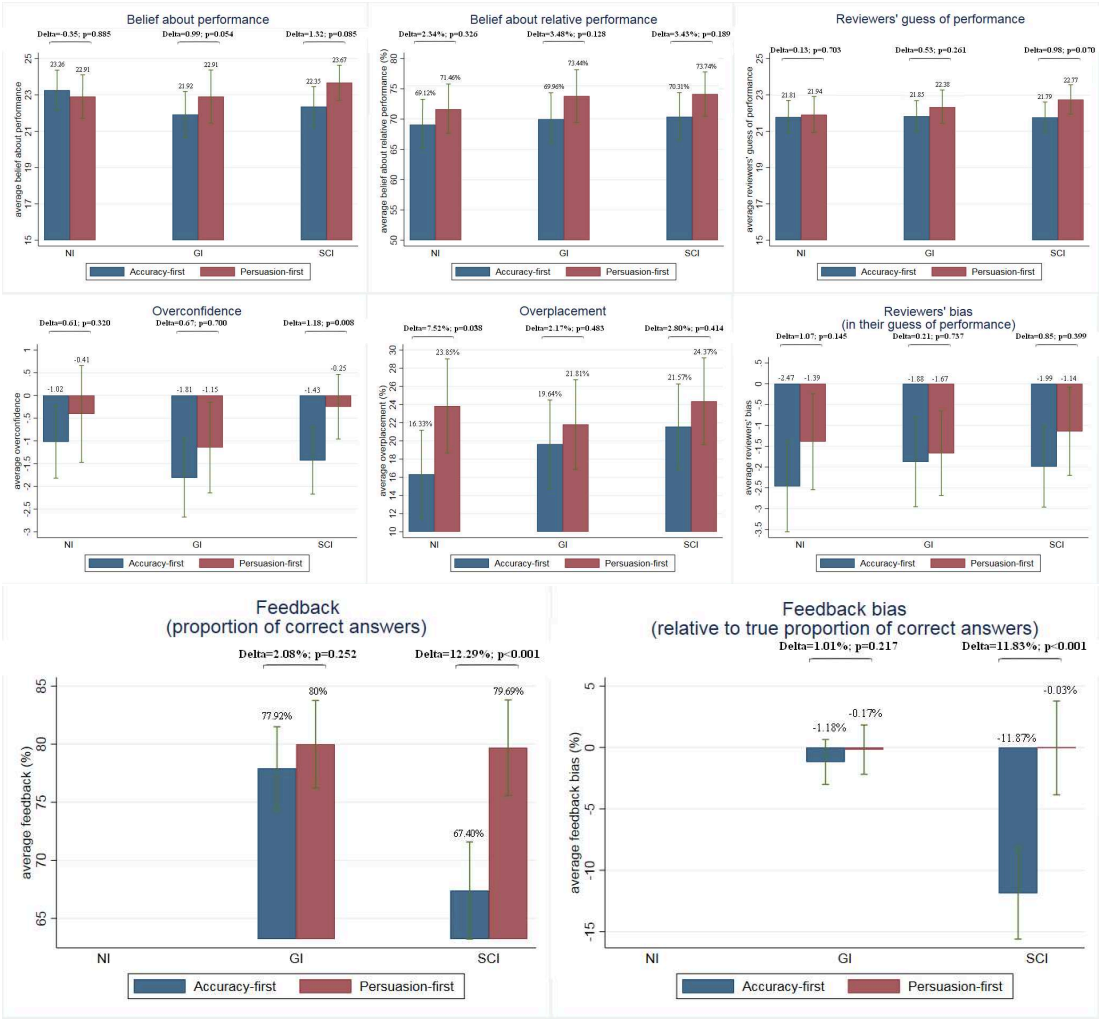


Figure 2.4: A summary of our main variables of interest comparing *Accuracy-first* treatment (left bars) and *Persuasion-first* treatment (right bars) across information conditions in Study 1 (with mean values, treatment differences and p-values (from two-sided Mann-Whitney rank-sum tests) indicated on top.

We also calculate a "*Feedback bias*" as the difference between the percentage of correct answers observed in the feedback questions and the percentage of correct answers received for the whole test. A positive feedback bias means that participants received feedback with a higher percentage of correct answers than the actual proportion of correct answers they achieved overall. If the sampling is unbiased, the expected proportion of correct answers revealed through the feedback should be equal to the actual proportion of correct answers. In that case, the "*Feedback bias*" will be equal to 0. As shown in bottom right panel in Figure 2.4, we find no significant difference in feedback bias in *Given Information*, as expected ($p = 0.217$). In contrast, we find a significant positive difference in

the *Self-Chosen Information* condition ($p < 0.001$). This difference indicates that participants in the *Persuasion-first* treatment chose to sample questions that they were more likely to have been right than those in the *Accuracy-first*, as suggested by Hypothesis 2.[25]

Result 2 (Biased information search) *Overall, participants in the Persuasion-first treatment sample more positive feedback than participants in the Accuracy-first treatment.*

In Hypothesis 3, we propose that in the situation with more mental wiggle room (i.e. *Self-Chosen Information*), participants will be able to form more confident beliefs. To test this hypothesis, we examine the treatment effect within each condition. Figure 2.4 presents the treatment comparisons on the same main variables of interest as in Figure 3 in the first two rows, but within each condition separately. We find that the participants' beliefs on performance and relative performance are almost always higher in the *Persuasion-first* treatment than in the *Accuracy-first* treatment (except beliefs about performance in *No Information* condition). However, most differences are not significant. Results on beliefs about their relative performances and the measure of overplacement also do not provide evidence for the effect of the treatments across different information conditions. When we use simple OLS regressions and pair-wise tests on the estimated treatment effects to directly test Hypothesis 3, we found no significant results.

Result 3 (Mental wiggle room & strategic confidence) *Inconsistent with Hypothesis 3, we find no clear evidence that the difference in participants' beliefs about their absolute and relative performances between Persuasion-first treatment and Accuracy-first treatment becomes significantly greater when they are given more freedom to select their feedback.*

Overall, Study 1 provides evidence for strategic use of overconfidence in social interactions (Hypothesis 1) and for biased information sampling (Hypothesis 2). Although Study 1 doesn't provide evidence for Hypothesis 3 and 4, it remains possible that the effect of the anticipation of strategic interactions may be stronger in the *Self-Chosen Information* condition, when participants can actively engage in selective information search.[26] Hence, it is possible that for self-deceptive overconfidence to emerge, sufficient mental wiggle room may be necessary. To assess

[25] An interesting finding in this study is that participants in *Accuracy-first* treatment (who are motivated to be as accurate as possible) sample 11.84% more difficult questions than those in the *Persuasion-first* treatment. Presumably feedback on the more difficult questions has greater probative value.

[26] This is suggested in the treatment comparisons on our primary measure of confidence and overconfidence in *Self-Chosen Information* condition (see the first two left panels in Figure 2.4).

this possibility, and to examine the robustness of our results, we report a replication of the *Self-Chosen Information* condition in the controlled environment of the laboratory in the next subsection. We then use our experimental results to identify the causal effect of information sampling on confidence and the causal effect of endogenously affected confidence on participants' persuasiveness, to assess the empirical links in von Hippel and Trivers (2011)'s theory.

2.3.2 Study 2: Replication in the Laboratory

Online experiments using MTurk appear to be reliable (Arechar et al., 2018), but the MTurk environment is not as controlled as in the lab. Moreover, the incentives on MTurk are quite low, which may reduce the motivation of the participants. Although we paid our participants on average more than twice the standard hourly rate typically available on MTurk, the reviewers were not incentivised to be accurate in their guess of the participants' score in Study 1. For these reasons, to ensure the reliability of the first study's results, we reproduce the *Self-Chosen Information* condition in a controlled laboratory environment in Study 2 at Queensland University of Technology (QUT). We recruited 100 QUT students for the main part of the experiment (50 in each treatment) and another 100 QUT students to participate as reviewers.[27] Participants and reviewers were both invited to the lab at the same time and separated into two different rooms on different floors upon arrival. At the end of the experiment, both the participants and the reviewers were paid in cash. The experiment was programmed using o-Tree (Chen et al., 2016). The experiment took on average 45 minutes and the average payoff was 11.20 AUD (s.e. = 3.58) for the main participants and 8.70 AUD (s.e. = 4.08) for the reviewers.

To implement the experiment in the laboratory we made some minor changes. First, we adapted 4 questions from the general knowledge test from Study 1 to make them more suitable for non-Americans participants. Second, each reviewer received four to six essays to ensure five independent assessments for each essay. Third, we incentivized the reviewers in the accuracy of their guesses about participants' score.[28] We did not provide incentives for the second question on

[27] We ran Study 2 with a smaller sample size than Study 1 as we expect our standard errors to be smaller in a more controlled environment.

[28] One essay was randomly drawn for each reviewer's payment according to the following rule: they would receive \$10 if their guess of the participant's score is equal to the participant's score or deviates from that score by only one item. They would receive \$8 if their guess deviates by two items. They would receive \$4 if their guess deviates by three items. They would receive \$2 if their guess deviates by four or five items and they would receive nothing if their guess deviates by more than five items. Note that participants were not informed about the incentives of the reviewers.

convincingness given that it is purely subjective.[29] Fourth, participants earnings were calculated (both for the Accuracy and Persuasion Task), by comparing them to the other participants in the same experimental session. This feature of the design allows us to pay participants at the end of each experimental session. Participants were aware of the reference group and reminded of the number of participants in their session before reporting their estimate. Finally, we asked participants to complete the 25-item version of the OCQ at the end of the experiment to avoid any impact OCQ might have on the main experiment.[30]

2.3.3 Results

Figure 2.5 displays the summary statistics for Study 2. The measures we use are exactly the same as in Figure 2.4.[31] We find that the results from Study 2 are generally very similar to those in Study 1. We observe no statistically significant difference in performances across treatments (20.02 *vs.* 18.92, $p = 0.200$), but the overall performance is slightly lower than in Study 1. In the *Persuasion-first* treatment, participants on average hold more favourable beliefs about their absolute performance (1.72 higher) and relative performance (6.76 percentage points higher). However, these differences are not significant (two-sided MW t-tests: p =0.178 and $p = 0.123$, respectively). When we examine biases in these beliefs, we find that the "*overconfidence*" measure is 2.82 higher in the *Persuasion-first* treatment (two-sided MW rank-sum tests: $p < 0.001$) but the "*overplacement*" measure remains nonsignificant ($p = 0.389$).

We also find that participants on average sample relatively easier questions in the *Persuasion-first* treatment and the proportion of correct answers (75%) is 12.8 percentage points higher than that in the *Accuracy-first* treatment ($p = 0.018$). The treatment comparison on the "*Feedback bias*" measure is also significant at 1% (11.93% vs. -4.53%; $p = 0.002$). In contrast to the *Self-Chosen Information* condition in Study 1 in which "Feedback bias" is mainly driven by participants sampling more difficult questions (-11.87%) in the *Acc.1st* treatment, the discrepancy in Study 2 is driven by deviations in opposite directions from both treatments. Namely, participants in the *Accuracy-first* treatment sample more

[29] Reviewers' ratings of convincingness are an auxiliary measure we use to ensure participants not only think about persuading reviewers that their performance is strong but also try to be as convincing as possible. Summary statistics provided in Table A.1 in the appendix show that there is no significant difference in reviewers' ratings of convincingess across treatments.

[30] Comparing Study 1 and 2, we find that it does not make a difference whether we place it at the beginning or the end of the experiment.

[31] A similar summary of the mean values and standard errors for all the variables is displayed in Table A.2 in the Appendix as well.

difficult questions (-4.53%) and participants in the *Persuasion-first* treatment sample easier questions (11.93%).

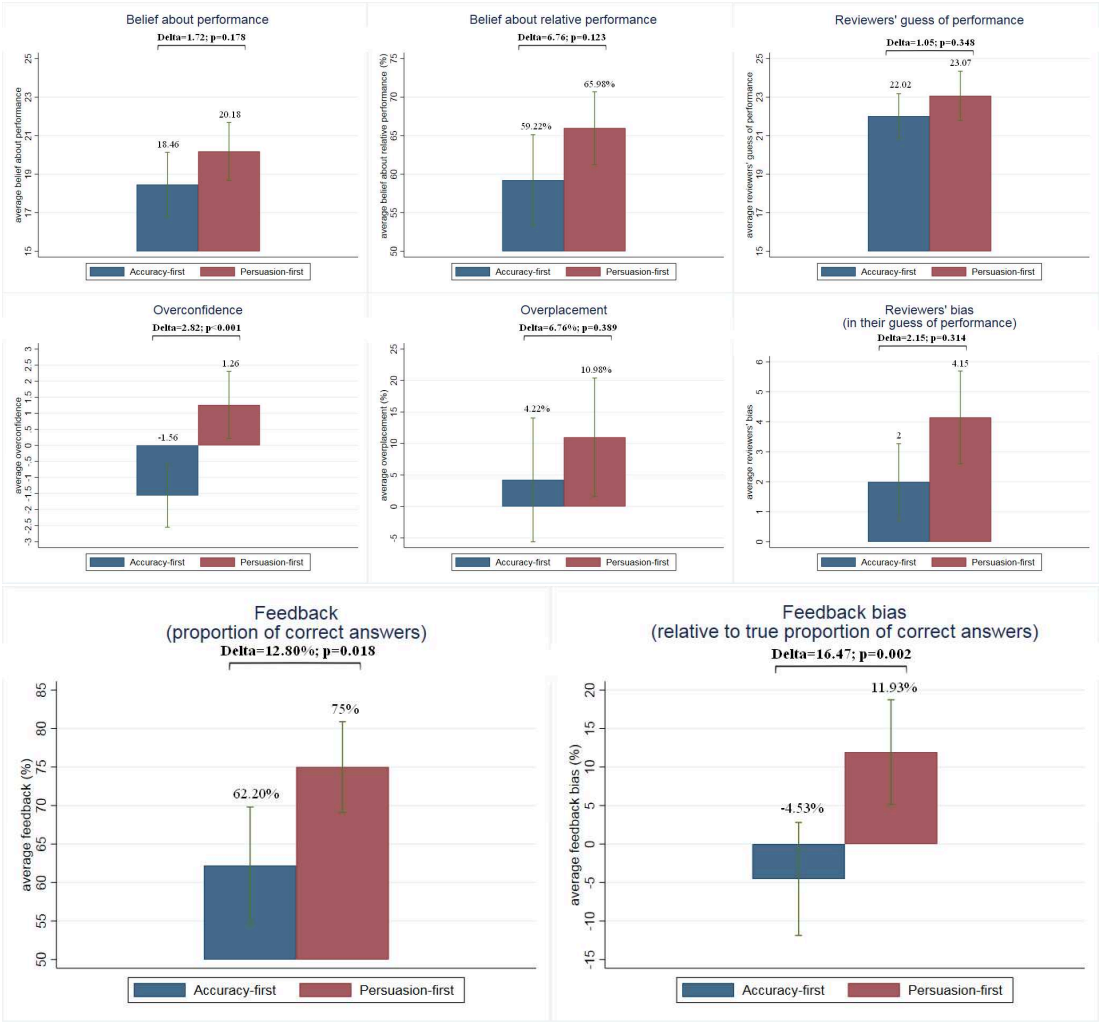


Figure 2.5: A summary of the mean values of our main variables of interest comparing Accuracy-first (blue bars) and Persuasion-first (red bar) in Study 2 (with mean differences between treatments and p-values indicated).

Finally, we find that the impact of the *Persuasion-first* treatment on the reviewers' estimates of participants' scores is also positive (1 item more), but not significant. The reviewers' bias in their guess of performances is also higher in *Persuasion-first* treatment (4.15 vs. 2), but the difference is again not significant ($p = 0.314$).

In summary, the additional results from Study 2 are largely consistent with the results of Study 1 and when taken together, these two studies provide evidence of both the existence of strategic overconfidence (Hypothesis 1) and selective information sampling (Hypothesis 2).

2.3.4 Causal Identification: Information Sampling on Confidence

As stated in Hypothesis 2, the observed differences in beliefs between participants in the *Accuracy-first* and *Persuasion-first* treatments in the *Self-Chosen Information* condition is likely to be facilitated by the tendency for people to bias their collection of information. Having established that participants in the *Persuasion-first* treatment indeed sampled information in a self-serving way, in this section, we further investigate how the bias in sampled information affects participants' beliefs.

Table 2.2: Causal identification of the effect of information sampling on beliefs about performance and relative performance.

Dep. Var:	*SCI* (MTurk)		*SCI* (lab)		*SCI* (MTurk +lab)	
Beliefs about	perf.	relative perf.	perf.	relative perf.	perf.	relative perf.
	(1)	(2)	(3)	(4)	(5)	(6)
Feedback	0.100**	0.259	0.644**	0.195**	0.132***	0.400***
	(0.047)	0.179)	(0.294)	(0.081)	(0.039)	(0.149)
Performance	0.672***	1.824***	0.703***	1.347**	0.670***	1.572***
	(0.128)	(0.491)	(0.158)	(0.570)	(0.087)	(0.329)
Constant	-0.356	9.484	-7.733	-7.789	-2.676	4.911
	(1.834)	(2.061)	(4.593)	(5.366)	(1.790)	(1.978)
First-stage F-stat	37.21	14.25	6.94	12.78	39.45	33.18
Observations	197	197	100	100	297	297

Notes: Column (1) to (6) report 2SLS regressions with standard errors in parentheses. *Feedback* is instrumented by the treatment dummy and is the *proportion* of correct answers contained in the sampled questions. Columns (1) and (2) shows the results from Study 1. Columns (3) and (4) shows the results from Study 2. Columns (5) and (6) shows the results for pooled observations from both studies. *** p <0.01, ** p <0.05.

If the hypothesis made by von Hippel and Trivers (2011) - that participants sample information in a self-serving manner to inflate their perceptions of their own performance - holds, we should observe that a higher proportion of correct answers in the feedback has a positive effect on participants' beliefs. In order to test this hypothesis, we could regress participants' beliefs about their performance on the feedback received (see Tables A.4 and A.5 in the Appendix). However, since participants are allowed to sample information at their own discretion, their feedback is likely to be endogenous. To address this issue, we instrument the "*Feedback*" variable by a dummy variable that equals 1 if the participant was in the *Persuasion-first* treatment and 0 if the participant was in the *Accuracy-first* treatment. This instrumentation requires the assumption that the *Persuasion-*

first treatment affects participants' beliefs only through the feedback received.[32] Table 2.2 reports the 2SLS regressions of participants' beliefs on "*Feedback*" (instrumented by the treatment dummy), controlled for actual performance. We use beliefs on absolute performance as the dependent variables in models (1), (3) and (5) and beliefs on relative performance in models (2), (4) and (6).[33] Columns (1) to (2) shows the results from Study 1. Columns (3) and (4) shows the results from Study 2. Columns (5) and (6) shows the results for pooled observations from both studies.

Models (1) to (6) show that an increase in the proportion of correct answers in the feedback increases participants' beliefs about their absolute and relative performance and the effect is significant at the 5% (1%) level for each individual study (pooled studies), with the exception of model (2). This result is consistent with a role of feedback on confidence. It is conditional on the identification assumption, and needs therefore to be interpreted with caution.

2.3.5 Causal Identification: the Effect of Confidence on Persuasiveness

The key hypothesis motivating our study is that overconfidence can arise strategically as people attempt to be more persuasive in social interactions. The above results provide evidence that when people anticipate a need to be persuasive, they form more favourable self-beliefs through biased information search. In this subsection, we estimate 2SLS regressions of participants' persuasiveness (measured by reviewers' average guessed scores) on participants' beliefs about their absolute and relative performance. To do so, we use the randomness of the treatment assignment to instrument participant's beliefs.

Our identification hypothesis is that participants' confidence is inflated through biased (self-serving) information search in the *Self-Chosen Information* condition and it is this inflated confidence which makes them more persuasive through written essays. This hypothesis essentially describes a channel through which confidence is enhanced from observing more positive feedback even though the positive feedback was endogenously manipulated. Given that participants in *Persuasion-first* treatment on average saw more positive feedback than those

[32] We find, in columns (9) to (14) from Tables A.4 and A.5 in the Appendix, that the treatment dummy is only significant when feedback is missing in the regression. Once the feedback is controlled, treatment dummy has no significant impact on beliefs, which suggests that the identification assumption for IV models is reasonable.

[33] Table A.7 in the Appendix also shows the same regressions with more control variables (sex, age, OCQ).

in the *Accuracy-first* treatment, another channel through which *Persuasion-first* treatment may affect reviewers' rating is directly through the content of written essays. For example, participants in the *Persuasion-first* treatment may be more likely to mention their feedback since they received more good news on average than participants in the *Accuracy-first* treatment. Mentioning the feedback itself may have a positive effect on reviewers' ratings. If the content of participants' essays differs systematically between treatments, our identification assumption may be violated. To assess this possibility, we recruited MTurk workers who never participated our experiment to read the essays and identified the type of messages written in the essays. Table A.10 in the Appendix shows that participants in the *Given Information* and *Self-Chosen Information* conditions in the MTurk study did mention their feedback more often in the *Persuasion-first* than in the *Accuracy-first* treatment. To control for this possible bias, we also add a "feedback dummy" in Table 2.3 that equals 1 if participants mentioned the feedback in their essays and 0 otherwise. Nevertheless, it is a restrictive assumption and our results need to be read in this light.

Table 2.3: Causal identification of the effect of participants' beliefs about performance and relative performance on persuasiveness.

Dep. Var:	*SCI* (MTurk)		*SCI* (lab)		*SCI* (MTurk + lab)	
Persuasiveness	(1)	(2)	(3)	(4)	(5)	(6)
Beliefs about perf.	0.866	–	0.508**	–	0.646***	–
	(0.618)		(0.254)		(0.213)	
Beliefs about	–	0.315	–	0.153**	–	0.210***
relative perf.		(0.271)		(0.063)		(0.074)
Feedback dummy	-0.423	-0.125	-0.646	-0.835	-0.236	-0.121
	(0.840)	(0.893)	(0.825)	(0.988)	(0.335)	(0.442)
Performance	-0.525	-0.510	-0.127	0.030	-0.364*	-0.261
	(0.558)	(0.661)	(0.207)	(0.147)	(0.195)	(0.193)
Constant	15.042***	11.836***	15.469***	12.713***	16.528***	13.800***
	(2.112)	(4.466)	(1.785)	(1.887)	(0.955)	(1.292)
First-stage F-stat	72.05	32.34	147.50	14.85	250.33	63.66
Observations	197	197	100	100	297	297

Notes: Table 2.3 reports 2SLS regressions for *SCI* conditions only with standard errors in parentheses. Participants' beliefs are instrumented by the treatment dummy. Columns (1) and (2) shows the results for observations from Study 1. Columns (3) and (4) shows the results for observations from Study 2. Columns (5) and (6) shows results for observations from pooled data. *** p <0.01, ** p <0.05.

In Table 2.3, we use beliefs regarding absolute performance in models (1), (3) and (5) and beliefs regarding relative performance in models (2), (4) and (6) as

the dependent variables. We also control for participants' performance in models (1) to (6). Columns (1) and (2) shows the results from Study 1. Columns (3) and (4) shows the results from Study 2. Columns (5) and (6) shows results from pooled data of Study 1 and 2.

Results from Table 2.3 show that participants' beliefs about their performance have a positive effect on the reviewers' average guess of participants' scores. These effects are not significant for Study 1 on both belief measures, but they are both significant at the 5% level in Study 2, where there is less noise and reviewers were incentivized by the correctness of their ratings. When pooling the data from the *Self-Chosen Information* condition of both studies to gain more power, the effect becomes significant at the 1% level in both models.[34] These results suggest that an increase in participants' confidence (particularly in study 2) may have a positive effect on their persuasiveness, even after controlling whether the feedback was mentioned to the reviewers.

Result 4 (Effectiveness of strategic confidence) *Participants holding higher self-beliefs (presumably generated via biased information sampling and motivated reasoning) tend to be more successful at persuading the reviewers that they did well in the test, particularly in the laboratory study.*

2.4 General Discussion and Conclusion

In the current research we tested the hypothesis that overconfidence emerges as a strategy to gain an advantage in social interactions. In service of this goal, we conducted two studies in which we manipulate participants' anticipation of strategic interactions and also the type of feedback they receive.

In our design, participants undertake both a Persuasion Task and an Accuracy Task in all treatments. By switching the order of these tasks, we can manipulate participants' goals (being accurate vs. persuasive). Because they were not aware of the nature of the second task when undertaking the first task, we prevent participants from engaging in a cost-benefit analysis between the two goals. However, we acknowledge that this choice of design has its own limitations.

First, self-deception might be possible in between the Accuracy and Persuasion Task in the *Accuracy-first* treatments. Because we did not elicit beliefs again

[34]Table A.8 in the Appendix displays similar (slightly weaker) results while adding more control variables (sex, age and OCQ).

after the Persuasion Task in the *Accuracy-first* treatments, we can not rule out this possibility directly. However, there is empirical evidence showing that the way people interpret information tends to be sticky. For example, Chambers and Reisberg (1985) presented participants with the famous duck/rabbit figure, which could be interpreted as either of these animals. They found that once participants arrived at an initial interpretation that it was a duck, they were unable to re-interpret it as a rabbit without seeing it again. In the same manner, our hypothesis was established on the expectation that once an (accurate) belief is formed, it is "on record". It can therefore not be consciously ignored by participants (even if they have incentives to form overconfident beliefs in the next task). Hence, without additional data, participants would not be able to re-construe their beliefs easily in our *Accuracy-first* treatment after the Accuracy Task.

In contrast, if a participant does not have a prior accurate belief "on record", it may be easier to interpret information in a self-serving manner. Similarly, we conjecture that once an inflated belief has been formed in the *Persuasion-first* treatment through motivated reasoning, it is also hard to "de-bias" it, even though the subsequent Accuracy Task required them to form the most accurate beliefs. There is no obvious reason to believe that participants were able to easily inflate beliefs (after forming well-calibrated beliefs in Accuracy Task) later in the Persuasion Task, but unable to easily deflate the overconfident beliefs (formed in the Persuasion Task) in the subsequent Accuracy Task. Our experimental results can be seen as justifying our assumptions ex-post, because we would have not found any treatment difference in belief elicitations if participants were able to adjust their beliefs flexibly depending on the incentives they were given in each task.

Second, the process of writing an essay in the Persuasion Task could lead participants to form inflated self-assessment of their performance, even in the absence of any self-deception motives. While there is evidence showing that self-introspection may lead to overconfident self-assessment (Wilson and LaFleur, 1995), Sedikides et al. (2007) find that written self-reflection actually decreases self-enhancement biases and increases accuracy.[35] If the writing task made it harder for the participants to form inflated beliefs, the treatment effect identified in the *Self-Chosen Information* condition might be underestimated. On the contrary if the writing task helped them form inflated beliefs, the effect size measured in the *Self-Chosen Information* condition might be overestimated. However, if the Persuasion Task itself inflated self-beliefs, we should have observed a significant treatment (*Persuasion-first* vs. *Accuracy-first*) difference in overconfidence in the *No Information* condition. The fact that we find no significant treatment

[35]See Zimmerman et al. (2011) for a similar example in the context of education.

difference on overconfidence in the *No Information* condition can be seen as tentative evidence that even if the Persuasion Task itself could inflate the beliefs, this effect is unlikely to be big enough to undermine the main effect we have identified in the *Self-Chosen Information* condition.

Finally, wishful thinking may play a role in our experiment. Apart from any strategic motives, participants may be willing to inflate their beliefs because they derive anticipatory utility from thinking that things will turn out well in the future (Mayraz, 2011; Heger and Papageorge, 2013). In the context of our experiment, participants may engage in wishful thinking because they believe that being of high performance leads to higher ratings, independently of the effect of beliefs on persuasion, but because reviewers may be able to detect high performers. Since participants in the Accuracy-first treatment are not aware that they will face the Persuasion Task, they have no incentive to engage in wishful thinking when they undertake the Accuracy Task. However, this is not the case for participants in the Persuasion-first treatment. Hence, treatment differences could be overestimated in our studies because our measures of confidence in the Persuasion-first treatment can reflect both wishful thinking and strategic motives. To disentangle these two motives, Schwardman and van der Weele (2019) elicited participants' beliefs about the role of ability on persuasion. The authors found no evidence that anticipatory utility drives their results and, more importantly, no evidence of an interaction between participants' persuasive goals and anticipatory utility. These findings suggest that wishful thinking is unlikely to be the main driver of our results.

Our findings from both studies support the idea that self-beliefs respond to variations in the incentives for overconfidence. In our experiments, participants were put in situations where they could receive higher payoffs from persuading other players that they performed well in a knowledge test. We observe that their confidence in their performance increased in such situations. Consistent with the interpretation that overconfidence is induced by strategic motivated reasoning, we observe that when given the freedom to choose their feedback, participants who were motivated to persuade chose to receive more positive information. This choice, in turn, helped them form more confident beliefs about their performance. Participants holding higher beliefs tend to be more successful at persuading the reviewers that they did well through a written essay, particularly in the laboratory study.

These results support the hypothesis that people tend to be more overconfident when they expect that confidence might lead to interpersonal gains, which helps

to explain why overconfidence is so prevalent despite the obvious costs of having miscalibrated beliefs. Future research should investigate whether the type of interpersonal advantage observed in the context of this experiment can also be observed in different strategic contexts (e.g. negotiation, competition).

Chapter 3

Strategic (Over)confidence in Negotiations[1]

3.1 Introduction

Overconfidence has been blamed as the main driver of the high rate of costly resolutions in a wide range of domains such as politics and foreign policies (Johnson, 2004). Hence, being able to accurately assess one's relative strengths plays an important role in avoiding violent resolutions. Nevertheless, individuals are often too confident about how much they deserve in the context of negotiations (Neale and Bazerman, 1983), leading to a high rate of violent resolutions that are costly for all sides. In the shade of these costs, it is unclear why this bias is so persistent in this context. While overconfidence can sometimes be attributed to computational errors or asymmetric information (Chambers and Windschitl, 2004), a growing body of evidence now suggests that this bias can also be motivated (Bénabou and Tirole, 2016). A specific strand of this literature posits overconfidence as an adaptive evolutionary strategy (Trivers, 1976) and suggests the existence of offsetting benefits that favor the emergence of this bias in social interactions (von Hippel and Trivers, 2011). In the present paper, we investigate (i) to what extent confidence affects the social outcome of a bilateral bargaining situation; and (ii) whether there exists some individual benefits that would rationalize the persistence of this bias in the context of negotiations.

To do so, we design an experiment in which we exogenously manipulate participants' beliefs about their relative performance. In the first part, participants undertake a 10-question general knowledge test in the form of a Multiple Choice

[1] Co-authored with Changxia Ke, Lionel Page and William von Hippel.

Questionnaire and are paid a piece-rate for each correct answer. Participants are then matched in pairs according to their performance at the test and a group account is allocated to each pair. In the second part, participants undertake a 30-question test similar to the one undertaken in the first part. Each correct answer by either of the participants in the same pair increases the value of the group account. Before beginning the second part, participants are informed that they will share their group account at the end of the test, without knowing the details of how the sharing will be decided. At the end of the second part, participants receive a noisy (unbiased) signal about their performance relative to their partner's. This manipulation generates an exogenous variation in participants' belief about their relative performance. The group account is then split in two unequal shares (70/30 percent of the group account) and pairs of participants have to agree on how to allocate the shares between them.

We use a 3-stage negotiation process in which pairs of participants are given an opportunity to reach an agreement in each of the 3 stages. In the first stage, participants claim the share they wish to receive and are asked to write a message to their partner to justify their choice. If they fail to agree on the allocation of the group account, participants enter a second stage in which they are given three additional minutes to reach an agreement. During these three minutes, they can communicate with their partner via an interactive chat. If they fail to reach an agreement within these three minutes, they enter a third stage in which they are given 30 additional seconds to reach an agreement. However, for each second spent in this stage, the shares decrease proportionally. If participants fail to reach an agreement before the end of the third stage, they both end up empty-handed.

Pre-existing experimental findings converge towards the conclusion that overconfidence is socially costly in the context of negotiations because it increases the risk of conflicting resolution. More confident negotiators have been shown to be more demanding (Kramer et al., 1993; McGillicuddy et al., 1984; Thompson and Loewenstein, 1992), which contribute to a higher rates of non-resolutions and, consequently, a waste of resources (Bazerman and Neale, 1982; Neale and Bazerman, 1985; Babcock et al., 1995a). In these studies, overconfidence has been shown to be negatively correlated with social efficiency. Because participants' beliefs are not exogenously manipulated in these experiments, causal inferences remain unsolved. Indeed, one cannot rule out potential sources of endogeneity such as reverse causality (success leads to confidence) or unobserved covariates such as beauty (Mobius and Rosenblat, 2006) or dispositional overconfidence (Paulhus et al., 2003).[2] Our paper contributes to this literature by showing the

[2]Dispositional overconfidence refers to the idiosyncratic level of overconfidence independent of

causal effect of an increase in confidence on the social outcome of a bilateral negotiation process.

This paper also relates to the literature on motivated beliefs, which posits that individuals strategically bias their beliefs when doing so leads to higher expected payoff (Bénabou and Tirole, 2016). A specific strand of this literature propose that overconfidence has considerable interpersonal value because overconfident individuals can advantageously influence others in social interactions (Heifetz et al., 2007; Johnson and Fowler, 2011; von Hippel and Trivers, 2011; Bénabou and Tirole, 2016). Recent experimental findings in economics support this idea. Charness et al. (2018) show that people were more likely to publicly state higher levels of confidence when doing so would deter a competitor to enter a tournament. In the experiments of Schwardman and van der Weele (2019) and Soldà et al. (2019), participants perform a task after which some of them are incentivized to convince others that they performed well. Half of the participants are informed about this opportunity to deceive others prior to privately stating their beliefs and the other half is not. Results from both experiments show that participants who expected to convince others that they performed well at the task formed positively biased beliefs which helped them to appear more convincing.

Findings from both empirical and theoretical work suggest that overconfidence may also emerge strategically in the context of negotiations. Babcock et al. (1995b) argue that holding positively biased belief about how much one deserves increases one's ability to advocate on behalf of one's own self-interest because *"it increases the sincerity behind attempts to persuade others and it makes it easier to take self-interested action while maintaining a belief that one is acting fairly."* (Swift and Moore, 2012, p. 272). Theoretical work from economics support this idea that overconfidence provides financial benefits in this context. Bar-Gill (2005) demonstrates that optimistic lawyers are more successful in extracting more favorable settlements. In line with this prediction, Kyle and Wang (1997) and Benos (1998) show that overconfident negotiators may generate higher expected gains compared to well-calibrated ones. Heifetz and Segev (2004) find that this holds only conditional on reaching an agreement. A few papers in psychology have focused on correlations between people's beliefs and their outcome from a negotiation. White and Neale (1994) and Galinsky and Mussweiler (2001) found a positive correlation between participants' reservation price and their claim in a buyer/seller setting. Moore (2004) found that under time pressure, the more people claim, the more they actually get, conditional on reaching an agreement. However, there is no causal evidence that more confidence leads to

any strategic motives.

higher expected payoffs in negotiation. This paper also contributes to this literature by providing causal evidence that being relatively more confident provides individual benefits in this context.

The closest study to ours is the empirical analysis conducted by van Dolder et al. (2015) on the data of the game show 'Divided', which inspired our design. In the game show, teams of three participants answer general knowledge questions and good answers are rewarded with money. At the end of the show, the money collected by the three participants is split in three unequal shares (roughly 60/30/10 percent of the jackpot). Participants have 100 seconds to agree on the allocation of the shares. However, for each second that passes, the shares decrease proportionally. If participants fail to agree on the allocation before the end of the allocated time, they all end up with nothing. While this natural experiment provides a close to ideal environment to study the role of overconfidence in negotiation, it is hard to infer causal relationship between participants' beliefs and the outcome of the negotiation as beliefs are not explicitly elicited in the game, nor can they be manipulated. Moreover, using data from a game show also raise selection issues as the participants were cast and chosen by the producers and decisions could be very different knowing the process will be broadcast publicly on TV.

Our results show that an increase in confidence at the pair level lowers the social outcome of the negotiation due to an increase in the occurrence of impasses and delays during the negotiation process. In contrast, we find that participants who are relatively more confident than their partner about their performance are more likely to end up with larger payoffs at the end of the negotiation process, which could explain the excess of confidence in negotiations. Our paper contributes to the literature in two ways. First, to the best of our knowledge, we provide the first causal evidence of the effect of individuals' beliefs on both the social and individual outcomes in the context of negotiations. Second, our results offer a rationale for the persistence of overconfidence in the context of negotiations by showing that being relatively more confident can be beneficial at the individual level.

The remaining sections are organized as follow. Section 3.2 describes the experimental design and our hypotheses. The data and results are presented in section 3.3. Section 3.4 concludes.

3.2 Experimental Design and Hypotheses

3.2.1 General Design

Our experiment is composed of 2 parts. In part I, participants are asked to answer 10 questions of general knowledge individually. For each question, they can choose the correct answer among four options. Participants receive 0.2 euro for each correct answer. At the end of part I, participants are ranked according to their performance at the task. The participant with the highest score is ranked 1 and the participant with the lowest score is ranked n (with n, the total number of participants in the session).[3] Participants are only informed about their payoff for this part at the end of the experiment.

In part II, participants are matched in pairs according to their rank: The participant ranked n is matched with the participant ranked $n-1$, the participant ranked $n-2$ is matched with the participant ranked $n-3$, and so on until all participants are matched. The matching procedure is common knowledge among the participants. They are then asked to answer 30 questions of general knowledge individually.[4] As in part I, the questions are the same for all participants and they can choose the correct answer among four options. Participants receive 0.67 euro for each correct answer. The money earned by both participants in the same pair is allocated to a group account. In order to prevent participants to infer their performance from the value of their group account (and ultimately their partner's performance), we added a random shock $e \in [-0.85; 1.15]$ on the productivity of the pair.[5] Let's denote p_i the number of correct answers of participant i and p_j the number of correct answers of participant j from the pair $\{i, j\}$. The value v of the group account of the pair $\{i, j\}$ is computed as follow: $v_{ij} = 0.67 * e * (p_i + p_j)$.

After participants have completed the 30 questions, we elicit their beliefs about their absolute and relative performance in part II. First, participants are asked to report their beliefs about the number of questions they answered correctly in part II. Participants receive 1 euro if their estimate is exact or deviates from their true performance by one question. They receive 0.50 euro if their estimate deviates from their true performance by more than one question but no more than two. If the estimate deviates by more than two questions, they do not earn nor lose anything. Then, participants are asked how likely they think they are to have

[3] Note that participants are not informed about their rank. However, they receive information about their score at the end of the experiment.

[4] The questions used in both parts of the experiment are displayed in Appendix B.5.

[5] For the same reason, we set the piece-rate in part II equals to a number with two decimal points.

outperformed their partner in part II. Participants indicate their belief on a scale from 0 to 100% on a slider without incentives.[6]

We then exogenously manipulate participants' belief about their relative performance by giving them a private (noisy) binary signal. Our procedure is similar to Schwardman and van der Weele (2019). Each participant faces two urns containing 20 balls of two different colors (red and green). The computer program selects a ball from one of these 2 urns. If the participant performed better than his partner in Part II, the ball is drawn from the urn with 15 green balls and 5 red balls. If the participant performed worse than his partner in part II, the ball is drawn from the urn with 5 green balls and 15 red balls. Therefore, a participant who outperformed his partner is more likely to see a green ball and a participant who was outperformed by his partner is more likely to see a red ball. We then elicit again participants' beliefs about their relative performance in part II. After the final belief elicitation, the value v_{ij} of the group account is displayed on the screen and participants are asked to decide how to allocate their group account via a 3-stage negotiation process.[7]

Negotiation: At the beginning of the negotiation process, participants are informed that their group account has been divided in two unequal shares. Their task is to reach an agreement on the allocation of these shares. The 'high' share is equal to 70% of the group account ($0.7v_{ij}$) and the 'low' share is equal to 30% of the group account ($0.3v_{ij}$). The negotiation process is divided in 3 stages displayed in Figure 3.1. Participants have the opportunity to reach agreement in each of the three stages. However, v_{ij} decreases in stage 3. The unfolding of the stages is described to the participants before they enter the negotiation process.

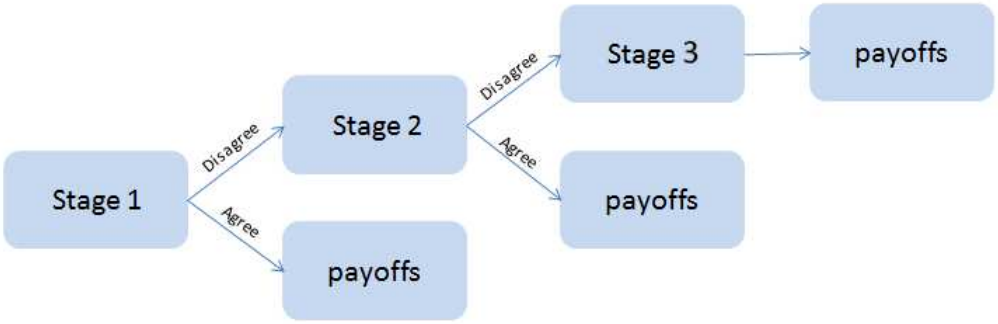


Figure 3.1: The three different stages of the negotiation process.

[6]The exact wording of the instructions is provided in Appendix B.6.

[7]Participants were told that they will have to split the group account at the beginning of Part II, but were only given instructions on the split and the negotiation procedure after the last belief elicitation.

In stage 1, participants are asked to claim either the high share or the low share and to write a message to their partner to justify their choice. There is no time constraint in this stage. If both negotiators from the same pair claim different shares, an agreement is reached: the participant who claimed the high share receives $0.7v_{ij}$ and the participant who claimed the low share receives $0.3v_{ij}$. In this case, the negotiation process ends in stage 1 and participants will not enter stage 2, nor stage 3. If both negotiators claim the high share in stage 1, they proceed to stage 2.[8]

In Stage 2, participants who did not agree in Stage 1 are given 3 additional minutes to try to reach an agreement. During these 3 minutes, participants can communicate via a chat box with their partner.[9] They are reminded of the amount allocated to each share, their own decision in stage 1 and their partner's decision in stage 1. They can decide to switch from the high share to the low share at any time by hitting the corresponding button on their screen. An agreement is reached when one of the negotiators in the pair switches from the high share to the low share. In this case, the participant who claimed the high share receives $0.7v_{ij}$ and the participant who claimed the low share receives $0.3v_{ij}$. The negotiation ends in stage 2 and participants will not enter stage 3. If no agreement is reached within the allocated time, participants proceed to stage 3.

In stage 3, participants are given 30 additional seconds to try to reach an agreement. However, for each second spent in this stage, the value of the shares decrease proportionally such that both shares will be equal to 0 at the end of the 30 seconds. Participants could observe on the screen the value of the shares decreasing in real time (i.e., shrinking every second). The shares stop shrinking when one participant chooses the low share. In this case, the participant who stuck to the high share receives the remaining amount allocated to the high share, and the participant who switched to the low share receives the remaining amount allocated to the low share. If no one switches before the end of the 30 seconds, both negotiators receive nothing and the total value of the group account is lost. The procedure for each stage of the negotiation process is described to the participants at the beginning of the negotiation phase. The unfolding of the experiment is displayed in Figure 3.2.

[8] Note that both participants would also enter stage 2 if they both choose the low share in stage 1. However, this situation never occurred in our data.

[9] The communication within pairs was only restricted in two ways: participants were not allowed to reveal the color of the ball that was shown to them and nor any private information that would uncover their anonymity.

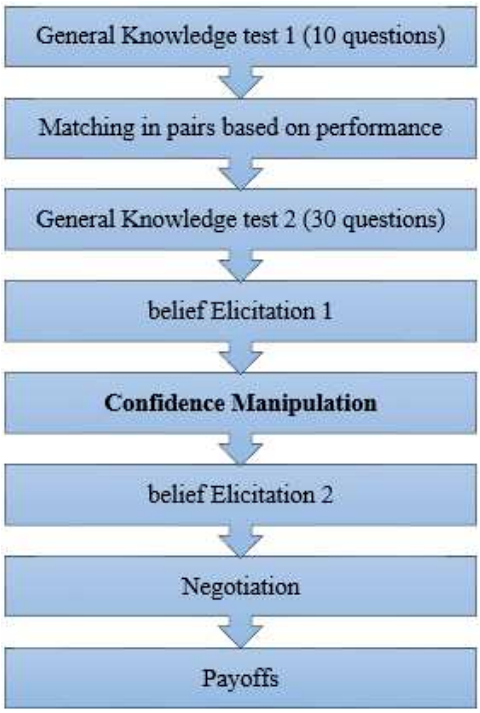


Figure 3.2: Unfolding of the experiment.

3.2.2 Hypotheses

In this experiment, pairs of two negotiators are asked to bargain over the unequal allocation of a prize whose value depends on their performance. The higher the performance of a negotiator relative to his partner, the more he contributed to the prize. Theoretical and empirical work show that individuals care about the proportionality between contributions and outcomes and deem fair that who contributed most to the prize should receive a larger share (Adams, 1965; Konow, 2003; van Dolder et al., 2015). If this is the case, participants who hold high belief about their performance relative to their partner's should feel "entitled" to a larger share of the prize and entitlement has been experimentally shown to influence bargaining behavior (Gächter and Riedl, 2005).[10] Hence, if both negotiators in the same pair are too confident that they deserve the larger share, conflicts are likely to arise.

In our setting, conflicting resolutions can take two forms: failure to reach an agreement and delays in reaching an agreement. Empirical evidence from bargaining experiments have shown that both are more likely to arise when both negotiators believe their side needs no compromise (Bazerman and Neale, 1982; Babcock et al., 1995a). This in in line with Heifetz and Segev (2004)'s evolutionary model of 'toughness' which predicts that when two tough negotiators meet,

[10]We define entitlement as how much one think he deserves based on one's contribution.

impasses and delays are likely. In addition, Ortner (2013) found that inefficient delays can arise when negotiators are too confident about their relative bargaining power. This leads to our two first hypotheses.

Hypothesis 1 *An increase in confidence at the pair level decreases the likelihood to reach an agreement.*

Hypothesis 2 *An increase in confidence at the pair level increases the duration of the negotiation process.*

By design, conflicting resolution are costly for both negotiators in the same pair. A failure to reach an agreement results in both negotiators leaving the negotiation empty-handed. Conditional on reaching an agreement, the value of the prize decreases for each second spent in the final stage of the negotiation process. Hence, delays in reaching an agreement can also be socially costly. This leads to the following hypothesis.

Hypothesis 3 *An increase in confidence at the pair level decreases the social outcome of the negotiation.*

Since being too confident is socially costly, Trivers (1976) argues there must be some offsetting material gains that would explain its persistence in such context. Bénabou and Tirole (2016) surveys evidence that individuals strategically manipulate their beliefs when doing so leads to higher expected payoffs. While most papers constituting this literature focuses on the intrapersonal advantages of motivated beliefs, a growing body of evidence from economics and psychology now shows that holding positively biased beliefs can lead to higher financial gain in strategic interactions (von Hippel and Trivers, 2011; Charness et al., 2018; Schwardman and van der Weele, 2019; Soldà et al., 2019).

Even though the existence of these strategic advantages has not been yet shown in the context of negotiations, early theoretical economic models predict that overconfident negotiators may generate higher expected gains (Kyle and Wang, 1997; Benos, 1998; Heifetz and Segev, 2004; Bar-Gill, 2005). Consistent with these predictions, some empirical evidence suggest a positive relationship between participants' beliefs about what they deserve and their outcome from the negotiation (White and Neale, 1994; Galinsky and Mussweiler, 2001; Moore, 2004). More confident negotiators have also been found to be more successful at convincing others that they deserve more (Babcock et al., 1995b). This leads to our final hypothesis.

Hypothesis 4 *The higher negotiators' beliefs are relative to their partners', the larger their payoffs are.*

3.2.3 Procedure

Our experimental design and hypotheses were pre-registered on AsPredicted.[11] We conducted this experiment at GATE-lab (Ecully, France). We recruited a total of 298 participants via Hroot (Bock et al., 2014), mainly among students from local engineering, business and medical schools. No subject participated in more than one session. We ran 21 sessions that involved an average of 14 participants per sessions. The experiment was programmed using o-Tree (Chen et al., 2016). Upon arrival, subjects were randomly allocated to a terminal. The terminal number was used as the participant ID for the payment collection. The instructions were distributed at the beginning of each part. The instructions for each part were read aloud by the experimenter. Participants were paid the sum of their earnings for each part in addition to a 5-euro show-up fee. The experiment took on average 1 hour (including payment) and the average payoff was 15.71 euros (s.e. = 0.389).[12] Participants received their payment in cash and in private at the end of the experiment.

3.3 Data and Results

3.3.1 Results on Beliefs

Figure 3.3 displays the average prior beliefs (light bars) and posterior beliefs (dark bars) for participants who received a bad signal and participants who received a good signal. The horizontal dashed lines represent the Bayesian posterior for a bad (-25) and a good (+25) signal, respectively. The bars of the histogram are between the two Bayesian updates, suggesting that participants are conservative on average (they update their belief conditional on their signal less than predicted by Bayes rule).

While there is no significant differences in prior beliefs conditional on the signal (two-sided Mann-Whitney test:[13] $p = 0.310$),[14] we found a strong significant difference in posterior beliefs depending on the signal received (MW test: $p < 0.001$). On average, participants update their beliefs in the direction of the signal they received: participants who received a bad signal update their beliefs downwards by 12.04 percentage points and participants who received a good signal

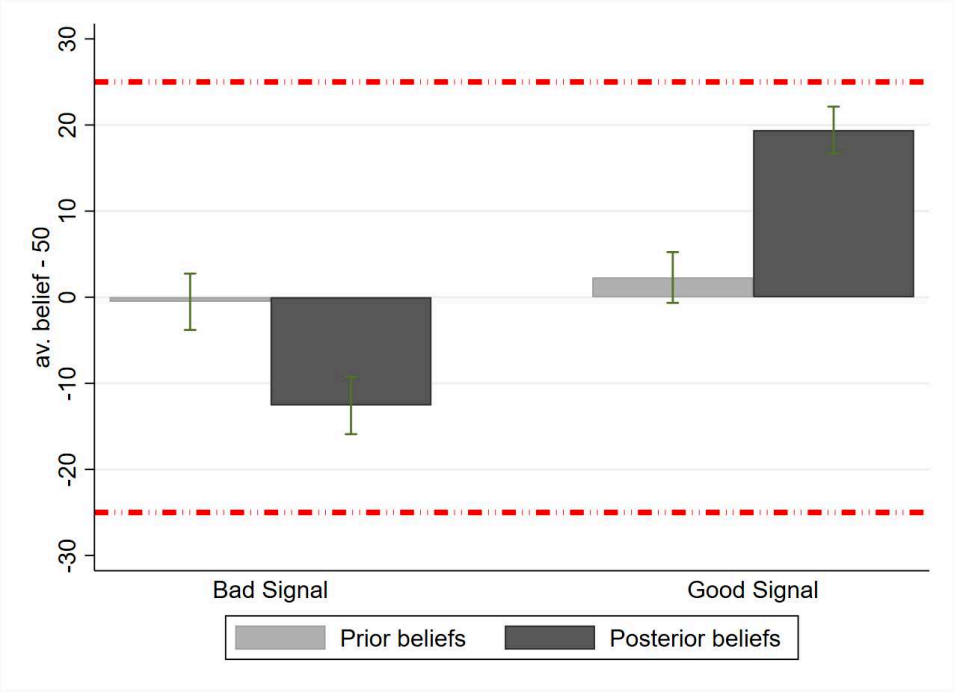
[11] The hypotheses statements were improved for exposition purposes. The pre-registration can be found at the following: https://aspredicted.org/bj9er.pdf.

[12] This includes the show up fee.

[13] MW test, hereafter.

[14] These priors are also not significantly different from 50% (Wilcoxon signed-ranks tests: p=0.820 for participants who received a bad signal and p=0.224 for participants who received a good signal)

update their beliefs upwards by 17.12 percentage points. This difference in updating suggests an asymmetry in the way our participants process good and bad news, which is supported by our analysis in Appendix B.3. These results show that our manipulation of participants' beliefs worked.



	B	G	Δ
prior beliefs (before the signal)	49.47%	52.29%	2.82%
	(1.656)	(1.493)	p=0.310
Posterior beliefs (after the signal)	37.43%	69.41%	31.97%
	(1.692)	(1.381)	$p < 0.001$
Observations	150	148	298

Note: We report the average prior and posterior beliefs for participants who received a bad (B) signal and participants who received a good (G) signal, as well as the p-values for two-sided Mann-Whitney tests between treatments (i.e., Δ). Standard errors in parentheses. The horizontal dashed lines represent the Bayesian posterior for a bad (-25) and a good (+25) signal.

Figure 3.3: Prior and posterior beliefs (normalized at 50%) about relative performance, by signal.

3.3.2 Confidence and Social Outcome

We hypothesize that high levels of confidence within a pair decreases the likelihood to reach an agreement (hypothesis 1) and increases the time spent in the negotiation process (hypothesis 2). In turn, we expect these conflicting resolutions to be socially costly (hypothesis 3). In section 3.3.1, we showed that participants who received a good signal form higher posterior beliefs than participants who

received a bad signal. Hence, the combination of signals received within a pair (two good signals (2G); two bad signals (2B); and two opposite signals (1G1B)) should be a good proxy of the level of confidence at the pair level. Table B.3 in Appendix supports this argument by showing that the higher the number of good signals in the pair, the higher the level of confidence at the pair level. We first examine the role of confidence on agreements failures and delays. We then investigate its effect on efficiency. Throughout the section we refer to Table 3.1 that displays summary statistics for agreements and efficiency, by combinations of signals.

Table 3.1: Summary statistics on agreements and efficiency, by combinations of signals.

	Agreements				**Efficiency**	
Combinations of Signals	stage 0 (1)	stage 1 (2)	stage 2 (3)	no agreement (4)	all (5)	all who agreed (6)
2G	0%	26.09%	47.83%	26.09%*	52.46%**	70.98%*
1G1B	6.86%	37.25%	44.12%	11.76%	80.57%	91.31%
2B	8.33%	41.67%	33.33%	16.67%	77.22%	92.67%
Obs.	9	54	64	22	149	127

Note: This table shows the proportion of participants who agreed in each stage of the negotiation process, the proportion of participants who failed to reach an agreement and the percentage of the initial group account left at the end at the end of the negotiation process. Stars indicate the results of two-sample tests of proportion and two-sample Mann-Whitney tests between pairs with two good signals and pairs with another combination of signals. *** p <0.01, ** p <0.05, * p <0.10.

Confidence, Agreements and Delays

In this section, we investigate the relationship between confidence and the likelihood that an agreement is reached, as well as the time spent in the negotiation process. Figure 3.4 displays the distribution of time needed to reach an agreement (in seconds). 6.04% of the pairs reach an agreement in stage 0, 36.24% in stage 1 (seconds 1 to 180) and 42.95% in stage 2 (seconds 181 to 220). 14.77% of teams did not reach an agreement at all and ended up empty-handed.[15] The spike around 180 seconds suggests that most people agree either at the end of the 3 minutes (14.77%) or immediately when the shares start to shrink (32.89%).

Columns (1) to (4) in Table 3.1 summarize the proportion of pairs of participants

[15]These numbers are surprisingly close from van Dolder et al. (2015) who found that in the Divided game show, 9% of the teams reach an agreement immediately, 72% later and 19% fail to do so.

who agreed in each stage of the negotiation process and the proportion of pairs of participants who failed to reach an agreement, conditional on the combination of signals they received. Two-sample tests of proportion show that there are marginally more pairs who failed to reach an agreement at any stage for pairs with two good signals than for pairs with one good signal (Two-sample tests of proportion: 26.09% vs. 11.76%; p =0.078).[16] The kaplan-Meier survival estimates provided in Figure B.2 in Appendix shows that this difference becomes significant at the 5% level when the sequential structure of the data is accounted for. Even though the proportion of agreements reached in stage 0 and 1 is also lower for pairs with two good signals than for pairs with a different combination of signals, the difference is not significant (PR tests: 2G vs. 1G1B, p =0.0.198 and p =0.0.313; 2G vs. 2B: p =0.0.196 and p =0.0.260). We do not find any significant difference in the subsequent stage. There is no significant difference between pairs with two bad signals and pairs with a different combination of signals. In addition, an analysis of the messages sent in Stage 1 in Appendix B.4 reveals that aggressive messages are negatively correlated with the likelihood to reach an agreement.

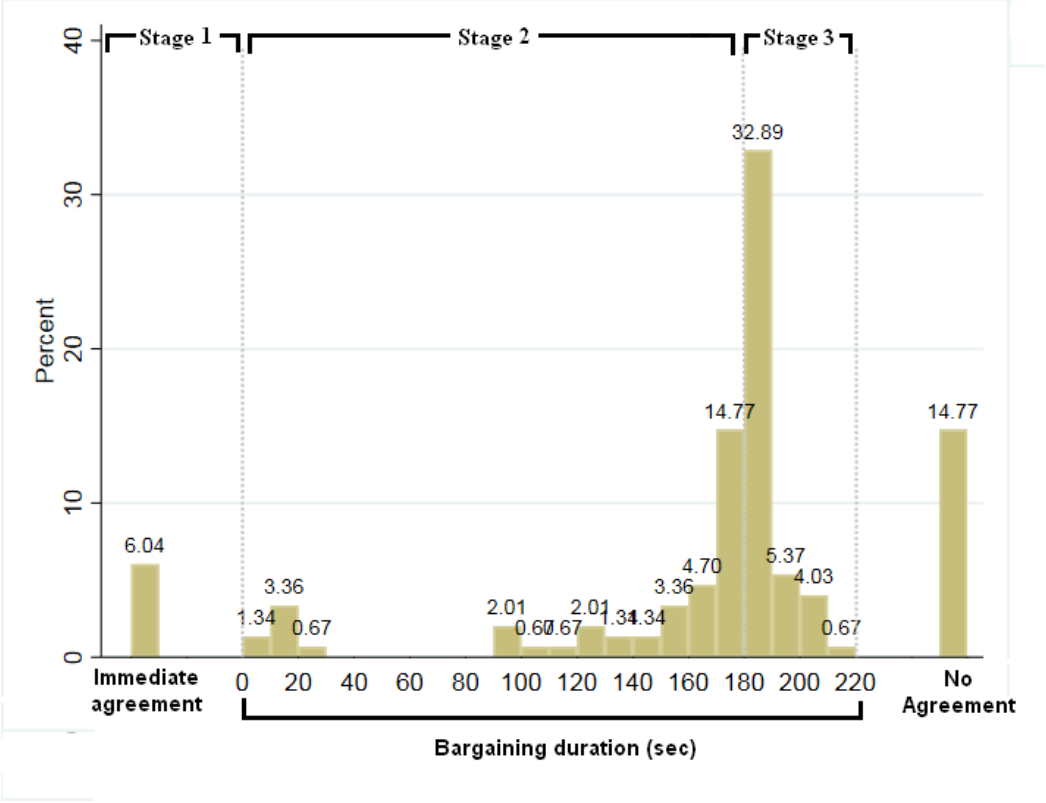


Figure 3.4: Summary statistics on agreement and efficiency.

We now investigate the causal effect of participants' beliefs on impasses and de-

[16]PR test, hereafter.

lays. Our data set provides information on whether pairs of participants reached an agreement or not, as well as the stage of the negotiation process in which an agreement occurred (if any). This data structure allows us to perform survival analysis to investigate the effect of confidence on the likelihood that an agreement is reached at a given point in time and, consequently, the time needed to reach an agreement. To do so, we estimate Cox regressions with proportional hazard in which the dependent variable is the rate of agreement (Columns (1) and (2) in Table 3.2).[17] The independent variables include the sum of beliefs of participants i and j from the pair $\{i, j\}$.

Table 3.2: Effect of confidence on delays and impasses.

Dep. var:	**Rate of Agreements** (t=stage)			
	(1)	(2)	(3)	(4)
$Belief_i + Belief_j$	-0.006**	-0.005*	-0.006*	-0.006*
	(0.003)	(0.003)	(0.004)	(0.004)
$score_i + score_j$	–	-0.015	–	-0.008
		(0.015)		(0.010)
Obs.	149	149	149	149

Note: Column (1) and (2) report the estimates of Cox regressions with proportional hazards of the sum of beliefs of participants i and j from the pair $\{i, j\}$ on the rate of agreements. Column (3) and (4) reports the GMM coefficients of Poisson regressions of the sum of beliefs of participants i and j from the pair $\{i, j\}$, instrumented by both i and j signals, on the rate of agreements. Standard errors in parentheses. *** p <0.01, ** p <0.05, * p <0.10.

However, participants beliefs may be correlated with unobservable characteristics. To rule out this potential source of endogeneity, we follow Schwardman and van der Weele (2019) and use the exogenous variation in confidence that results from the noise component of the feedback signal as an instrument. We showed in section 3.3.1 that the signal shifts participants' beliefs because it is informative about the true state of the world (whether the participant performed better than his partner in part II). However, conditional on the true state of the world, the signal is completely random and exogenous, making the signal a good candidate for an instrument. In addition, participants were instructed not to disclose their signal in their messages. The signal therefore can only influence outcomes through its impact on private beliefs. This guarantees the validity of our instrument. To implement this procedure, we estimate GMM Poisson regressions in

[17] The survival functions in Figure B.2 in Appendix suggests that the effect of confidence on the rate of agreements is not time-dependent. The assumption of proportional hazard is therefore reasonable.

which the dependent variables is the same as in the Cox regressions (Columns (3) and (4) in Table 3.2). The independent variables include the sum of beliefs of participants i and j from the pair $\{i, j\}$, instrumented by both i and j signals.[18] We also control for the sum of participants i and j performance in part II in models (2) and (4). Our unit of observation is one pair.

Models (1) to (4) show that an increase in confidence at the pair level decreases the likelihood to reach an agreement and, consequently, increases the time spent in the negotiation process. The results are significant at the 5% level in model (1) but only marginally significant in models (2) to (4). However, one question that naturally follows is to what extent these two undesirable effects are detrimental for efficiency. This is the focus of the next section.

Result 1 *An increase in confidence at the pair level decreases the likelihood to reach an agreement (supports Hypothesis 1).*

Result 2 *An increase in confidence at the pair level leads to more delays during the negotiation process. (supports Hypothesis 2).*

Confidence and Efficiency

The efficiency rate in our experiment (measured as the fraction of the group account that is actually awarded) is 76.02%, meaning that almost 24% of the overall value created during the effort task is wasted in delays and impasses. Columns (5) and (6) of Table 3.1 show the percentage of the initial pot left at the end of the negotiation process for all pairs and pairs who reached an agreement only, conditional on the combination of signals received by the pairs of participants. Two-sample Mann-Whitney tests show that the percentage of the group account left at the end of the negotiation process is significantly smaller for pairs of participants who received two good signals than for pairs of participants who received two opposite signals, and this result holds weakly (due to a smaller number of observations) when considering only pairs of participants who reached an agreement (MW tests: $p = 0.017$ and $p = 0.083$, respectively). There is no significant difference between pairs with two bad signals and pairs with a different combination of signals.

To investigate the causal effect of participants' beliefs on efficiency, we estimate OLS regressions in which the dependent variable is the fraction of the initial group account that is left at the end of the negotiation process. The independent

[18] Note that without the instrumentation, GMM Poisson regressions lead to the exact same results as Cox regressions with proportional hazard.

variables include confidence at the pair level, measured as the sum of beliefs of participants i and j from the pair $\{i, j\}$, instrumented by both i and j signals. We also control for the sum of participants i and j performance in Part 2. Our unit of observations is one pair. The results are reported in Table 3.3. Model (1) shows the results for all pairs. Model (2) shows the results for pairs of participants who reached an agreement only.

Table 3.3: Effect of confidence on the social outcome of the negotiation.

Dep. var: Fraction awarded of the group account	all (1)	agreements only (2)
$Belief_i + Belief_j$	-0.004**	-0.003***
	(0.002)	(0.001)
$score_i + score_j$	0.002	0.006
	(0.005)	(0.003)
Constant	1.121***	1.037***
	(0.213)	(0.137)
Obs.	149	127

Table 3.4: *Note:* Table 3.3 reports the OLS estimates of the sum of beliefs of participants i and j from the pair $\{i, j\}$, instrumented by both i and j signals on the fraction awarded of the initial group account. Column (1) reports the results for all pairs. Column (2) shows the results for pairs of participants who reached an agreement only. Standard errors in parentheses. *** p <0.01, ** p <0.05, * p <0.10.

Model (1) in Table 3.3 shows that an increase in confidence leads to a decrease in the final fraction of the initial group account that is awarded and the results are significant at the 5% level. Model (2) shows that this result holds when considering only pairs of negotiators who reached an agreement before the end of the negotiation process and the results are significant at the 1% level.

All together, these findings provide evidence that being too confident leads to delays and impasses in negotiations and that these two phenomena are significantly detrimental for the social outcome of the negotiation process.

Result 3 *An increase in confidence at the pair level leads to a lower social outcome of the negotiation (supports Hypothesis 3).*

3.3.3 Confidence and Individual Outcome

Our main hypothesis is that the most confident negotiators of the pair will gain an advantage in the negotiation process from being more confident than his partner (hypothesis 3). If this is the case, participants who received a good signal

should end up with a larger payoff from the negotiation than their partner, when their partner received a bad signal.

Table 3.5: Outcome of the negotiation process, by signals.

Combination of signals	Obs.	Payoff (in AUD)	Payoff (% group account)	% participants with high share
B_iB_j	46	9.64	38.61%	41.67%
		(1.019)	(0.037)	(0.072)
G_iB_j	102	10.79**	45.17%***	57.84%***
		(0.658)	(0.025)	(0.049)
B_iG_j	102	8.71	35.39%	30.39%
		(0.592)	(0.021)	(0.046)
G_iG_j	48	6.39	26.23%	36.96%
		(1.018)	(0.039)	(0.072)

Note: Table 3.5 displays the average payoff in both AUD and percentage from the pot, as well as the percentage of participants who ended up with the high share at the end of the negotiation process. Stars indicate Wilcoxon matched-pairs signed-ranks tests comparing payoffs from the negotiation (AUD and % group account) and two-sample tests of proportion comparing the proportion of participants who end up with the high share, comparing G_iB_j and B_iG_j. * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

Table 3.5 displays participants' payoff from the negotiation both in AUD and in percentage of the initial group account, and the proportion of participants who end up with the high share, conditional on both the signal they received and the signal their partner received. Hence, for each pair of participants $\{i, j\}$ we have four possible combinations of signals: i received a bad signal while j received a good signal (B_iG_j); both i and j received a bad signal (B_iB_j); both i and j received a good signal (G_iG_j) and i received a good signal while j received a bad signal (G_iB_j).

Results from Table 3.5 shows that in pairs of participants who received two opposite signals, the participant who received the good signal receives on average 45.17% of the group account (10.79 AUD) while his partner receives on average 35.39% of the group account (8.71 AUD) and the differences in both payoff measures are significant (Wilcoxon matched-pairs signed-ranks tests: $p = 0.018$ and $p = 0.001$, respectively). We also find that 57.84% of participants who received the good signal end up with the high share while only 30.39% of participants who received the bad signal end with the high share and the difference is significant at the 1% level (PR test: $p < 0.001$). Table B.5 in Appendix shows that participants who received a bad signal are more likely to switch from the high share to the

low share than participants who received a good signal, which is consistent with the findings from Table 3.5.

Table 3.6: Effect of relative beliefs on participants' payoff from the negotiation.

Dep. var:	All		Agreements only	
% of group account	(1)	(2)	(3)	(4)
$Belief_i - Belief_j$	0.002**	0.002**	0.002**	0.002**
	(0.001)	(0.001)	(0.001)	(0.001)
$Score_i - Score_j$	-0.002	-0.002	-0.003	-0.003
	(0.006)	(0.006)	(0.005)	(0.005)
Age		-0.001		-0.003
		(0.003)		(0.004)
Female		-0.054		-0.009
		(0.041)		(0.037)
Risk preferences		0.041		-0.003
		(0.011)		(0.010)
Constant	0.397***	0.432***	0.440***	0.541***
	(0.021)	(0.114)	(0.018)	(0.110)
Obs.	149	149	127	127

Note: Table 3.6 shows the results of the 2SLS estimations of the percentage of the group account received at the end of the negotiation on difference between participants' beliefs within the same pair, instrumented by the signals received by both participants. Standard errors in parentheses. *** p <0.01, ** p <0.05, * p <0.10.

To investigate the causal effect of relative confidence on the outcome of the negotiation, we estimate 2SLS regressions in which the dependent variable is the percentage of the group account received at the end of negotiation process. The independent variables include the difference in posterior beliefs between participant i and participant j from the same pair $\{i, j\}$, instrumented by a dummy variable equals to 1 if participant i received a good signal, and 0 otherwise; and a dummy variable equals to 1 if participant j received a good signal, and 0 otherwise. We also control for the actual difference between participants i and j scores. We only consider one participant per pair. The estimates are displayed in Table 3.6. Models (1) and (2) show the results for all pairs. Models (3) and (4) shows the results for participants who reached an agreement only. In models (2) and (4), we control for participants demographics (sex, age and risk preferences).

Models (1) to (4) in Table 3.6 show that the more confident a negotiator is relative to his partner, the more he receives from the group account at the end of

the negotiation process and the effect is significant at the 5% level even when considering pairs of participants who did not reach an agreement. Interestingly, results from columns (1) and (2) show that the benefits from being more confident remains, even when considering pairs that did not reach an agreement.

In summary, this section highlights the duality of overconfidence in a bilateral negotiation setting: on the one hand, being the most confident negotiator in the pair provides a significant financial advantage, which provides an incentive for negotiators to become more confident. On the other hand, if both negotiators are overconfident, money is wasted in costly delays and both partners end up worse off.

Result 4 *Negotiators who are more confident than their partner end up with a larger payoff from the negotiation (supports Hypothesis 3).*

3.4 General Discussion and Conclusion

Overconfidence has been identified as the main driver of costly delays and impasses in bargaining situations and yet, this bias often arises in this context. If overconfidence has actual material costs in negotiations, an adequate explanation of its persistence from an evolutionary point of view most likely requires for it to provide some offsetting benefits as well.

To examine this trade-off, we designed a laboratory experiment based on a 3-stage bilateral negotiation process. In this experiment, participants are matched in pairs and asked to undertake a general knowledge test individually. Each correct answer provided by either of the participants in the same pair earns money that is allocated to a group account. At the end of the task, we manipulate participants beliefs about their performance relative to their partner's by using a private noisy binary signal about their performance. Participants then have to agree on how to allocate their group account through a 3-stage negotiation process. However, the allocation is constrained to a 70%-30% split. If participants do not reach an agreement after three minutes, the value of their group account decreases such as both participants end up empty ended after 30 seconds.

We find that an increase in confidence at the pair level increases the likelihood of impasses and the time spent in the negotiation process, which - consequently - lowers the social outcome of the negotiation. In addition, we find that these inefficiencies may be driven by the stubbornness of participants with high level of confidence who are less likely to give up on the largest share. Finally, we

show that an increase in relative confidence leads to larger payoffs at the end of the negotiation process: negotiators who are relatively more confident than their partner end up with a larger share of the initial group account. These gains generate a demand for inflated beliefs which, in turn, increase the likelihood of costly resolutions. Overall, our findings are in line with the idea that people manipulate their beliefs strategically when doing so can lead to higher expected payoffs.

We acknowledge some limitations of our experiment. In particular, our experimental design does not allow to separate a strategic increase in confidence from wishful thinking. Apart from any strategic motives, participants may be willing to inflate their beliefs because they derive anticipatory utility from thinking that things will turn out well in the future (Mayraz, 2011; Heger and Papageorge, 2013). In the context of our experiment, participants may engage in wishful thinking if they expect that being of high type will lead to more favorable outcome. This assumes that most participants can detect high performance and deem to be fair that high performing participants receive a larger share of the group account. Since wishful thinking can be a potential confound, we cannot rule out the possibility that our results are upward biased (because wishful thinking would also lead to inflated beliefs). However, an excess in confidence is particularly costly in our setting, especially for pairs with two good signals who lose on average about half of the total value of their group account in the negotiation process. From an evolutionary perspective, it is unlikely that overconfidence would arise to provide only psychological benefits while having actual material costs (Trivers, 1976; von Hippel and Trivers, 2011). Hence our results are unlikely to be driven by wishful thinking.

Our findings contribute to the literature by providing a rational for the persistence of overconfidence in the context of bilateral negotiation despite its social cost. If an increase in confidence can lead to higher individual gains in negotiations, there is no reason to expect agents to be well-calibrated in such context. Our experiment also highlights the role of information in bargaining protocols. Despite being quite conservative in their beliefs updating, participants who received a positive signal exhibited higher levels of confidence than their counterparts who received a negative signal. This increase in confidence was, in turn, sufficient to provide an advantage during the negotiation process. Hence, these effects are expected to be even bigger in real-world situations with larger stakes, generating incentives to engage in self-serving strategies such as biased information search and information avoidance. Future research should investigate how current bargaining protocols can be improved to prevent/mitigate the detrimental impact of these incentives (e.g. transparency, third-party arbitration, etc.).

Chapter 4

Overconfidence as a Strategy in Leadership Striving[1]

4.1 Introduction

Self-confidence plays an important role in the determination of leaders (Shamir et al., 1993) because it is a quality people favors in their leaders (Hogan et al., 1994). People are attracted to leaders that display cues of confidence such as low pitch voices (Klofstad et al., 2012; Tigue et al., 2012) and high height (Blaker et al., 2013; Stulp et al., 2013) and can recognize leaders by their facial features with above-chance accuracy (Todorov et al., 2015). In turn, confident leaders are perceived as more knowledgeable (Price and Stone, 2004) as well as more trustworthy (Penrod and Cutler, 1995), and exert greater influence (Van Swol and Sniezek, 2005). While these results suggest that self-confidence helps in the process of reaching higher social status, evidence suggest that overconfidence play a similar role. In a theoretical model, Goel and Thakor (2008) show that, under certain conditions, overconfident managers have a higher probability to be promoted than well-calibrated managers. Reuben et al. (2012) find that the under-representation of female as leaders in a competitive setting was mainly driven by differences in overconfidence between men and women, suggesting that people favor overconfident leaders.

Most of the time, high social statuses provide access to scarce resources (Berger et al., 1972; Blau, 1964; Griskevicius et al., 2010; Savin-Williams, 1979; Ellis, 1994; Keltner et al., 2003). A growing body of evidence now support the idea that overconfidence is an effective strategy to attain such resources (Bénabou and

[1]Co-authored with Changxia Ke, Lionel Page and William von Hippel.

Tirole, 2016). In the context of social interactions, overconfidence has been shown to enable overconfident individuals to advantageously influence others (Heifetz et al., 2007; Johnson and Fowler, 2011; von Hippel and Trivers, 2011). Charness et al. (2018) show that participants are more likely to over-report their beliefs about their strengths when it is optimal to deter opponents' entry into a contest. Furthermore, Schwardman and van der Weele (2019) and Soldà et al. (2019) find that overconfident individuals are more successful in their attempt to persuade others that they performed well in a task.

In this paper, we investigate whether overconfidence emerges as a strategy in leadership striving when the leader's position provides privileged access to monetary benefits. To do so, we design a laboratory experiment in which we exogenously manipulate the incentives to be a group leader across treatments. After completing an effort task, participants are matched in groups of four and asked to select a leader for the group. Participants communicate with their group members via an online chat-box and then vote privately and simultaneously to select a leader. The leader will then make a series of binary choices that affect the payoffs of everyone in the group. The likelihood to receive the highest payoff from these binary choices depends on whether the leader is the best performer in the group or not. In the Symmetric Incentives (SI) treatment, the leader's payoff is the same as the rest of the group. In the Asymmetric Incentives (AI) treatment, the leader receives an additional payment. Our design allows us to investigate (i) how varying incentives may affect participants' (over)confidence which may further distort the selection of a group leader; and (ii) how leaders' confidence influences their decisions, and consequently the group welfare.

The closest study to ours is Anderson et al. (2012) who investigate the causal effect of overconfidence on social status. The authors design a series of experiments in which participants are matched in pairs and perform a task together. At the end of the task, each participant is asked to rate their partner on several measures that will define their social status.[2] The authors found that individuals who are overconfident about their performance appear competent to others and in turn, reach higher social status. In these experiments, social status is self-reported and has no direct consequence on the partner's payoff. Moreover, the payoff structure of both partner is always symmetric (i.e. there are no incentives to reach a higher social status in the pair except a preference for social status *per se*). Our experiment differs from Anderson et al. (2012) in two ways. First, we compare a situation similar to the one illustrated in Anderson et al. (2012), with a situation

[2]Measures include the degree to which the partner deserved respect and admiration, had influence over the decisions, led the decision-making process, and contributed to the decisions.

in which there is an extra financial incentives to reach a higher social status. Second, the decisions made by the high status group member affect the social welfare.

Our paper broadly relates to the literature on leadership. Experimental data show that the existence of leaders has a positive effect on contribution to a Public Good Game (PGG) and coordination on the desirable outcome.[3] However, the role of overconfidence in the selection of these leaders has not been at the center of the investigation. In this strand of the literature, the role of leader is determined by random assignment (Brandts and Cooper, 2007; Gächter et al., 2012; Drouvelis and Nosenzo, 2013; Gächter and Renner, 2014; Boulu-Reshef et al., 2015; Brandts et al., 2016), being the first-mover (Brandts et al., 2007; Rivas and Sutter, 2011), volunteering [4] (Haigner and Wakolbinger, 2010; Arbak and Villeval, 2013), or performance (Frackenpohl et al., 2016). When the leader is selected by the group, it is either based on past contribution to the PGG (Güth et al., 2007; Levati et al., 2007; Hamman et al., 2011; Markussen and Tyran, 2017) or on attributes unrelated to the tasks (Brandts and Cooper, 2007; Levy et al., 2011; Brandts et al., 2015). These experiments tend to show that the existence of leaders is socially beneficial. However, in these settings, the leaders' decision has little to no consequence for the group outcome. When it does, the competence of the leader does not affect the outcomes of his decisions. Moreover, the payoff function of the leader is the same as any other group member.

Our paper also relates to the literature on overconfidence and risk. Findings from economics have shown that overconfidence affects risk attitudes. Barron and Gravert (2018) design an experiment in which they manipulate participants' confidence about their performance at a task. Participants are then asked to choose their payment scheme between a risky lottery and a fixed piece-rate. The authors found that participants choose the risky incentives more often after an increase in confidence. Barber and Odean (2001) show empirically that overconfident male traders take inconsiderate risks. In the context of leadership, Heaton (2002) and Malmendier and Tate (2005) found that overconfident CEOs make poor investments or mergers decisions. Consistent with these findings, the theoretical model by Goel and Thakor (2008) predicts that an overconfident manager chooses higher levels of project risks when they are competing for leadership.

[3] See Güth et al. (2007); Arbak and Villeval (2013); Boulu-Reshef et al. (2015); Markussen and Tyran (2017); Levati et al. (2007); Haigner and Wakolbinger (2010); Hamman et al. (2011); Levy et al. (2011); Rivas and Sutter (2011); Gächter et al. (2012); Drouvelis and Nosenzo (2013); Gächter and Renner (2014); Jack and Recalde (2015); Brandts et al. (2016); Frackenpohl et al. (2016) for PGG experiments and Gillet et al. (2011); Brandts et al. (2015); Brandts and Cooper (2007); Brandts et al. (2007) for coordination games.

[4] Group members indicate their wish to lead. If more than one participant wishes to lead, the leader is selected randomly among those participants).

However, these papers focuses on the role of overconfidence on leaders' decisions but not on its role on the selection of these leaders.

We differ from the existing literature in three ways. First, we provide a situation in which the payoff function of the leaders is different from the payoff function of the other group members. This feature allows us to investigate whether overconfidence emerges in situations that offers additional gains for being the leader and is closer to real-life situations. Second, the leader's decisions has actual consequences on the other group members payoff, which allows us to study the social consequences of overconfidence.[5] Finally, the outcome of the leader's decisions is determined by the leader's ability.[6] By giving participants the clear incentive to select the high ability group member as the leader, we create a direct social cost of overconfidence.

Our findings show that participants who hold higher beliefs being the top-ranked performer in their group also have higher chances to be selected as the leader of their group. However, the top-ranked group member is almost twice as likely to be chosen as the leader in the SI treatment (58.6% of the time) than when the leader receives an additional fixed amount for being the leader (32.3% of the time). These unqualified leaders make overconfident decisions that lead to lower payoffs for their group members compared to leaders who do not receive such bonus. These findings suggest that aspirants for the leadership who expect to be rewarded by a bonus (i) are less likely to be the top-ranked member in their group and (ii) make overconfident choices that earn less money for their group than leaders who did not receive a bonus. These findings highlight the downside of monetary incentives: while aiming to reward competence, monetary incentives lead to the emergence of unqualified leaders when overconfidence is perceived as actual competence.

The remaining of the paper is organized as follows: Section 4.2 details the experimental design and hypotheses. In Section 4.3, we display our main results and describes the procedure. In section 4.4, we provide a discussion on our results and conclude.

[5] In previous experiments, the leader show the example by being the first to contribute to the PGG (Güth et al., 2007; Gächter et al., 2012; Arbak and Villeval, 2013; Frackenpohl et al., 2016; Levati et al., 2007; Haigner and Wakolbinger, 2010; Rivas and Sutter, 2011; Drouvelis and Nosenzo, 2013; Gächter and Renner, 2014; Jack and Recalde, 2015; Brandts et al., 2016) or send a message to the group (Levy et al., 2011; Boulu-Reshef et al., 2015). In Güth et al. (2007) and Levati et al. (2007), the leader can also exclude group members from participating in the public good in some treatments.

[6] In Hamman et al. (2011) and Markussen and Tyran (2017), the leader makes the contribution decision in a PGG for the whole group but the leader's decision does not depend on the leader's ability.

4.2 Experimental Design and Hypotheses

4.2.1 General Design

This experiment is composed of two parts. In part I, participants have to solve 20 Raven's matrices individually. They have 1 minute per matrix. For each matrix completed correctly, participants receive a piece-rate w. Therefore, the payoff function for an individual i in part 1 is the following: $\pi_{1i} = w$*score$_{1i}$. Participants will only learn about their payoff at the end of the experiment.

At the end of the task, each participant is ranked from 1 to N according to their performance (the participant with the highest score is ranked 1 and the participant with the lowest score is ranked N).[7] Participants are then split into 4 quartiles based on their rankings. Groups of 4 participants are then formed in the following way: each group is composed of one participant randomly drawn from each of the four quartiles.[8]. In that way, each group has one (and only one) group member who is ranked in the top 25% of participants in the room and this is common knowledge.

Participants are told that they will have to select a leader and that the leader will make a series of decisions that will affect the payoffs of everyone in the group. Each group member is more likely to receive more money from the leader's decisions if the leader is the top-ranked member in the group based on performances in Part 1. Participants do not know the exact nature of the decisions that the leader has to make, but they know their payoff function and they know that it is in their best interest to elect the group member who has the highest rank in their group.[9] To further insure that group members are incentivized to select the best performer in their group, they also receive \$0.5 for each matrix correctly solved by the leader in Part 1 in addition to the payment from the leader's decision.

In summary, participants' payoff in part 2 is decomposed into two parts: A part from the leader's performance: as in part 1, each participant (leaders and followers) receives a piece rate w (\$0.50 in our experiment) for each task completed

[7] N can be 8, 12 or 16 depending on the session. Tied performances are ranked randomly within their corresponding ranks.

[8] Namely, one participant is from the top 25%; one participant is from the top 50% (but not in the top 25%); one participant is from the bottom 25% and one participant is from the bottom 50% (but not from the bottom 25%)

[9] Revealing the nature of the leader's decisions may lead participants to vote according to their risk preferences rather than their beliefs about participants performance in part 1. In order to avoid this confounding motivation, we decided to limit the information provided on the leader's decisions before a leader is selected.

correctly by the leader in part 1 (score_l). A part from the leader's decision: each participant (leaders and followers) receive the payoffs p_l generated by the leader's decision.

In the Symmetric Incentives (SI) treatment, the payoff function of both leaders and followers is the same. Therefore, the final payoffs π_{2i} of a participant is determined by:

$$\pi_{2i} = w * score_{1,l} + p_l \tag{4.1}$$

In the Asymmetric Incentives (AI) treatment, the payoff function of the followers is the same as in the SI treatment but the leader receives an additional bonus B for being the leader on top of other payoffs. Therefore, the payoffs π_{2i} of a participant in the AI treatment is determined by:

$$\pi_{2i} = \begin{cases} w * score_{1,l} + p_l & \text{if the participant is a follower} \\ w * score_{1,l} + p_l + B & \text{if the participant is a leader} \end{cases} \tag{4.2}$$

Before Part 2 starts, participants are first asked to answer questions about instructions for Part 2. And then they are asked to state their beliefs about: (1) the number of matrices they solved correctly in Part 1 (from 0 to 20); (2) the average number of matrices correctly solved by the other participants in the room in Part 1 (from 0 to 20), (3) their beliefs about their rank in the session (a number from 1 to N). We use incentive compatible mechanisms to reward participants for being accurate on these three estimates. Finally, we also ask participants to state their beliefs about the probability that they are the best performer in their group. Participants then enter a public chat where they can communicate with the other members of their group for at most 10 minutes.[10] After the chat, participants are asked again to report their belief about the probability that they are the best performer in their group and make their choice of leader privately and simultaneously.[11] The leader is the group member who received the most votes. In case of ties, the computer program randomly allocate one extra vote to one of the participants with the highest number of votes.

After a leader is selected, participants' roles (either Leader or Follower) are displayed on their screen. Participants assigned to the role of leader undertake a

[10] Participants were given the opportunity to end the chat sooner if they wanted to by pressing a button on the screen.

[11] Participants are allowed to vote for themselves.

series of 10 binary choices.[12] Leaders were allowed to make multiple switches. Leaders' choices are displayed in Table 4.1.

Table 4.1: Leader's decision

Decision	Option A	Option B
1	\$10 if you are the best performer in your group; \$0 if not	\$1
2	\$10 if you are the best performer in your group; \$0 if not	\$2
3	\$10 if you are the best performer in your group; \$0 if not	\$3
4	\$10 if you are the best performer in your group; \$0 if not	\$4
5	\$10 if you are the best performer in your group; \$0 if not	\$5
6	\$10 if you are the best performer in your group; \$0 if not	\$6
7	\$10 if you are the best performer in your group; \$0 if not	\$7
8	\$10 if you are the best performer in your group; \$0 if not	\$8
9	\$10 if you are the best performer in your group; \$0 if not	\$9
10	\$10 if you are the best performer in your group; \$0 if not	\$10

Table 4.2: Sequence of the experiment

	Symmetric Incentives (SI)	**Asymmetric Incentives (AI)**
PART 1	Effort task ranking and group assignment Information on Incentives (symmetric)	Effort task ranking and group assignment Information on Incentives (asymmetric)
PART 2	Belief elicitations 1, 2, 3 and 4 Interactive Chat Belief elicitation 5 Leader's selection Leader's decision	Belief elicitations 1, 2, 3 and 4 Interactive Chat Belief elicitation 5 Leader's selection Leader's decision
	Payoff	Payoff

For each of the 10 decisions, leaders have to choose between Option A and Option B. While Option A always stays the same, the amount in Option B increases incrementally. A leader who believes that he has a 60% chance to be the best performer in the group has an expected payoff from Option A of \$6. Therefore, this participant should choose option A for Decision 1 to 5, be indifferent between option A and option B for decision 6 and finally choose option B for Decision 7 to 10. At the end of the experiment, one of the leader's decisions is randomly

[12]We use the same binary choices as in Barron and Gravert (2018).

chosen for payment and determines the payoff of everyone in the group for this part.

Finally, participants undertake the 'Assertiveness' scale from the Achievement Motivation Scale (Cassidy and Lynn, 1989) that allows us to measure participants' social dominance.[13] We also elicit participants' sex, age, risk preferences and whether English is their native language or not. Table 4.2 summarizes the structure of this design.

4.2.2 Hypotheses

A growing body of experimental evidence now supports this idea that individuals use their beliefs about their performance to influence others in strategic interactions (Anderson et al., 2012; Charness et al., 2018; Schwardman and van der Weele, 2019; Soldà et al., 2019). Therefore, the incentive to become overconfident is expected to be modulated by the existence of possible gains from doing so. In the Asymmetric Incentives treatment, leaders are given a bonus of 5 AUD. Hence, this treatment gives participants an incentive to successfully convince their group member that they are the top-ranked performer in the group. This leads to our first hypothesis.

Hypothesis 1 *Participants hold higher beliefs about their relative performance in the Asymmetric Incentives (AI) treatment than in the Symmetric Incentive (SI) treatment, which will facilitate their efforts to be chosen as leaders.*

When the incentives of group members are the same, individual self-interest is in line with the group interest: every group member has an incentive to select the top-ranked performer as their group leader. On the contrary, asymmetric incentives create a conflict between the group interest (choosing the top-ranked performer as the leader) and individual self-interest (earning the bonus). Hence, even though some participants will still put in their best effort in trying to identify the best performer in the group, some others will try to convince others that they are the best performer in the group even when they are not. As a consequence, it is harder for participants in the Asymmetric Incentives (AI) treatment to identify and elect the true top-ranked performer as a leader. This leads to the following hypothesis.

[13] Burks et al. (2013) find that socially "dominant" individuals exhibit more overconfidence. They conjecture that this is because socially dominant individuals attribute more importance to the belief of others about their ability.

Hypothesis 2 *The Asymmetric Incentives (AI) treatment will lead to the emergence of less qualified leaders compared to the Symmetric Incentives (SI) treatment.*

Once leaders have been chosen, they have to make a series of decisions. For each decision, they can either choose a fixed amount of money or a lottery that will earn them 10 AUD if they are the best performer of their group and 0 if they are not. If overconfident leaders (after being elected) truly believe they are more likely to be the best performer in their group, they will favor risky to safe options more often than they should. In line with this idea, overconfidence has been shown to affect risk attitudes (Barber and Odean, 2001; Barron and Gravert, 2018): overconfident CEOs make poor investments decisions (Heaton, 2002; Malmendier and Tate, 2005) and overconfident manager chooses higher levels of project risks when they are competing for leadership Goel and Thakor (2008). This leads to our final hypothesis.

Hypothesis 3 *Unqualified and overconfident leaders who emerge in the Asymmetric Incentives (AI) treatment will make sub-optimal decisions that will impair the group's welfare.*

4.2.3 Procedures

This experiment was pre-registered on AsPredicted.[14] We conducted this experiment at Queensland University of Technology (in Brisbane, Australia) in 2018. We recruited a total of 240 QUT students via ORSEE. No one participated in more than one session. We ran 21 sessions with either 16, 12 or 8 participants per sessions. We collected observations for 31 groups in the Asymmetric Incentives treatment and 29 groups in the Symmetric Incentives treatment. The experiment was programmed using o-Tree (Chen et al., 2016). Upon arrival, subjects were randomly allocated to a computer terminal. The terminal number was used as the participant ID for final payment collection. The instructions for the preliminary part were distributed at the beginning of the session and the instructions for the subsequent parts were distributed at the beginning of each part. After reading the instructions, the experimenter goes through a summary of the key points of the instructions using a powerpoint presentation. Participants then complete a quiz to check their understanding of the instructions for Part 2. Participants were paid the sum of their earnings for each part in addition to a $5 show-up fee. The experiment lasted on average 1 hour (including payment) and the average payoff

[14]The pre-registration can be found here: https://aspredicted.org/d8cz4.pdf.

was \$25.04 (s.e. = 0.301).[15] Participants received their payment in cash and in private at the end of the experiment.

4.3 Data and Results

4.3.1 Results on Beliefs

Treatment and Confidence

Table 4.3: Mean values and standard errors of main measures.

	SI	AI	Δ	*p*-values
Overconfidence (1)	1.97	1.98	0.01	0.723
	(0.250)	(0.276)	(0.374)	
(leaders only)	1.55	2.16	0.61	0.255
	(0.420)	(0.466)	(0.630)	
Overplacement (by % outperformed) (2)	12.55%	11.33%	-1.22%	0.823
	(2.653)	(2.576)	(3.698)	
(leaders only)	4.60%	12.71%	8.11%	0.176
	(5.891)	(5.507)	(8.057)	
Proportion of overplacement (by quartile) (3)	47.41%	45.16%	-2.25%	0.727
	(0.047)	(0.045)	(0.065)	
(leaders only)	34.48%	45.16%	10.68%	0.403
	(0.090)	(0.091)	(0.128)	
Estimation bias (4)	0.62	0.66	0.04	1.000
	(0.222)	(0.235)	(0.324)	
(leaders only)	0.72	0.71	-0.01	0.994
	(0.385)	(0.407)	(0.562)	
Belief about proba. to be in top 25% (before chat) (5)	57.06	56.81	0.25	0.939
	(2.309)	(2.250)	(3.225)	
(leaders only)	68.83	66.61	2.21	0.749
	(4.846)	(4.884)	(3.417)	
Belief about proba. to be in top 25% (after chat) (6)	56.76	52.99	3.77	0.268
	(2.399)	(2.395)	(3.393)	
(leaders only)	77.41	68.03	9.38	0.088
	(3.280)	(4.231)	(5.404)	
N	116	124	240	

Note: Table 4.3 shows the mean for overconfidence, overplacement, accuracy bias and the probability to be the top-ranked member in the group before and after the chat. We display the values first for all participants and then for leaders only. We report the p-values of two-sample Mann-Whitney tests between treatments for measure (1), (2), (4), (5) and (6); and p-values of two-sample test of proportion for measure (3) (i.e., Δ). Standard errors in parentheses.

Table 4.3 displays the mean values (with standard errors in parentheses) of (1) participants' overconfidence, measured as the difference between participants' be-

[15] This includes the show up fee. All amounts are in Australian dollars.

liefs about their score in Part 1 and participants' actual score; (2) participants' overplacement, measured as the difference between participants' beliefs about the percentage of participants they outperformed in Part 1 and the actual percentage of participants they outperformed;[16] and (3) the proportion of participant who overplace themselves by at least 1 quartile. We also measure (4) the 'estimation bias' as the difference between participants' belief about the average score in Part 1 of all participants in their session and the actual average score of the session. Finally, we elicit participants' belief about their probability to be in the top 25% of the distribution of performance (5) before and (6) after the chat. For each measure, we first display the mean values of all participants in each treatment, followed by a summary for leaders only.[17]

Table 4.3 shows that in both treatments, participants are overconfident regarding their performance (+1.97 in the SI treatment and +1.98 in the AI treatment) and this is significantly different from the zero-mean error (Wilcoxon signed-rank tests: $p < 0.001$, for both treatments).[18] Participants also overestimate the average performance of others in the session, in both treatments (+0.62 in the SI treatment and +0.66 in the AI treatment) and this is also significantly different from the zero-mean error (W tests: $p = 0.008$ and $p = 0.013$, respectively). Moreover, participants overestimate themselves more than they overestimate others, in both treatments (1.97 vs. 0.62 in the SI treatment and 1.98 vs. 0.66 in the AI treatment) and this is significant (W tests: $p < 0.001$ for both treatments). These differences shows that overconfidence is not purely driven by an estimation bias. Participants also show overplacement in both treatments (+12.55% in the SI treatment and +11.33% in the AI treatment) and this difference is significantly different from the zero-mean error (W tests: $p < 0.001$ for both treatments). We also find a substantial proportion of participants who overplaced themselves by at least one quartile (47.41% in the SI treatment and 45.16% in the AI treatment). We also found that participants are overconfident about their probability to be the best performer in their group, both before and after the chat in both treatments (compared to the average theoretical proportion of 25%). However, we do not find any significant differences in any of our measures of overconfidence and overplacement between treatments. Results displayed in Table C.4 in Appendix also show no causal effect of the treatment on any of the self-reported measures of confidence.

[16] Given n, the number of participants in the session, these beliefs are converted from participants' belief about their absolute rank B_r, and is calculated as follow: $(n\text{-}B_r)/n^*100$. The total number of participants varies across sessions (8, or 12, or 16), so does the number of possible ranks. To pull all sessions in a treatment together, we need this conversion.

[17] See participants' raw beliefs in Table C.2 in the Appendix A.

[18] W test, hereafter.

Looking at leaders only, leaders in the AI treatment overestimate their absolute performance by 50% more than participants in the SI treatment (+2.16 vs. +1.55), while there is no difference in accuracy bias (+0.71 vs. +0.72) but this difference is not significant.[19] The difference in overplacement is 8.11pp between both treatments (4.60% vs. 12.71%) but the difference is not significant. Leaders' beliefs about their probability to be the best performer in their group are higher after the chat (compared to their beliefs before the chat) in both treatment (68.83% vs. 77.41% in the SI treatment and 66.61% vs. 68.03% in the AI treatment) but this difference is only significant in the SI treatment (W test: $p = 0.022$ and $p = 0.547$, respectively). We provide an analysis of the effect of the chat on beliefs in Appendix C.3.

Confidence and leadership

In order to investigate the effect of confidence on leaders' selection, we estimate Poisson regressions in which the dependent variable is the number of votes received by participant i. The independent variables include participant's belief about his or her probability to be the top-ranked member of the group. Results are reported in Table 4.4. Models (1) and (2) consider all the observations. Models (3) and (4) consider the observations for the SI treatment only. Models (5) and (6) consider the observations for the AI treatment only. We control for the percentage of participants outperformed in models (1), (3) and (5), and participants demographics (sex, age, risk preferences and social dominance) as well as session size fixed effects in models (2), (4) and (6). Table 4.4 shows a positive relationship between belief about the probability to be the best performer in the group and the number of votes received by participant i. These results suggest that individuals who hold higher beliefs also receive the most votes. Table C.5 in Appendix displays the same pattern when using participant's belief about the percentage of participants he outperformed as the main regressor.

We hypothesized that participants in the AI treatment will become overconfident in order to be more successful at convincing others that they are the top-ranked member of their group and, in turn, increase their likelihood to receive the leader's bonus. We find that participants are significantly overconfident in both treatments. In addition, we found that participants who hold higher beliefs about their relative performance also receive the most votes. However, we do not observe any significant treatment effect on beliefs. This result fail to provide

[19] As shown in Table C.2, this difference is mainly driven by a lower performance of leaders in the AI treatment. We will discuss the leaders selection later.

sufficient evidence to support Hypothesis 1. We will further elaborate what might have driven this result later in the discussion section.

Table 4.4: Determinants of votes.

Dep. variable:	pooled		SI		AI	
Nb. of votes received	(1)	(2)	(3)	(4)	(5)	(6)
Belief top-ranked	0.018***	0.016***	0.017***	0.018**	0.018**	0.015*
	(0.005)	(0.006)	(0.007)	(0.007)	(0.007)	(0.009)
percentile	0.012***	0.011***	0.019***	0.020***	0.005	0.005
	(0.004)	(0.004)	(0.006)	(0.006)	(0.005)	(0.006)
N=8		Ref.		Ref.		Ref.
N=12		0.020		0.039		0.002
		(0.083)		(0.147)		(0.131)
N=16		-0.086		-0.209		-0.072
		(0.088)		(0.199)		(0.127)
AMS score		0.037*		0.026		0.049*
		(0.022)		(0.031)		(0.029)
Female		-0.092		0.073		-0.231
		(0.169)		(0.263)		(0.256)
Age		-0.003		0.000		-0.003
		(0.016)		(0.025)		(0.020)
Risk preferences		-0.044		-0.024		-0.065
		(0.042)		(0.055)		(0.061)
Constant	-1.727***	-2.119***	-2.151***	-2.709**	-1.391***	-1.743*
	(0.350)	(0.718)	(0.478)	(1.070)	(0.491)	(1.049)
Obs.	240	240	116	116	124	124

Note: Table 4.4 shows the results of the Poisson estimation of the number of votes received by participant i on participant i's belief about the probability that he is the top-ranked member of his group. Standard errors in parentheses. *** p <0.01, ** p <0.05, * p <0.10.

Result 1 *Participants hold high beliefs about their performance, which facilitate their efforts to be chosen as leaders. However, we find no evidence that this effect is driven by the treatment (partially supports Hypothesis 1).*

4.3.2 Results on Leaders' Selection

In the SI treatment, payoffs of both leaders and followers are aligned. Hence, participants in the SI treatment should do their best to identify the group member who performed the best. On the contrary, the AI treatment creates an asymmetry in payoffs between leaders and followers. Therefore, we should expect participants in the SI treatment to identify the best performer in their group more often than participants in the AI treatment. Table 4.5 shows the proportion of leaders who (1) are the top-ranked member; and (2) obtained the best score in their group,

across treatments.[20] We can see that in the SI treatment, participants select the top-ranked group member as the group leader almost twice as often as in the AI treatment and more than twice as often if we look at the best score, and these difference are significant (two-sample tests of proportion: $p = 0.040$ for top-ranked leaders and $p = 0.010$ for leaders having the best scores, respectively).[21]

Table 4.5: Distribution of leaders.

Proportion of leaders:	SI	AI	Δ	p-value
who are top-ranked (1)	58.62%	32.26%	-26.36%	0.040
	(0.091)	(0.084)	(0.124)	
with the best score in their group (2)	65.52%	32.26%	-33.26%	0.010
	(0.088)	(0.084)	(0.122)	
N	29	31	60	

Note: Table 4.5 shows the average proportion of leaders ranked in the top 25% and the average proportion of leaders who obtained the best score in their group, for both treatments. We report the p-values of two-sample tests of proportion between treatments (i.e. Δ). Standard errors in parentheses.

We found that in the AI treatment, leaders are less likely to be the top performer of their group than in SI. One question that can arise is whether each group member is equally likely to be selected as the leader or whether participants from a specific quartile are more likely to be selected than participants from the top quartile. Table C.6 in Appendix supports our previous findings that participants in the top 25% are more likely to be selected as leader in the SI treatment than in the AI treatment, and the effect is significant at the 1% level. In addition, our results show that in the AI treatment, participants from the second quartile (in the top 50% but not in the top 25%) are more likely to be selected as leader than participants in the first quartile and the effect is significant at the 5% level.[22]

In summary, participants who belong to the top 25% are more likely to be chosen as leader in the SI treatment. On the contrary, participants who belong to the top 50% but not in the top 25% are more likely to be selected as leader in the AI treatment than participants who belong to the top 25%.

[20] If several participants obtained the same score, the computer program randomly break the ties by allocating each participant to a unique rank. Hence, it is possible that a participant who obtained the highest score in his group belongs to the second quartile instead of the first one. For the sake of consistency, we will drop these observations when analyzing the leaders' decisions in section 4.3.3.

[21] One can think that this difference may be driven by participants in the AI treatment voting more often for themselves than participants in the SI treatment. Appendix C.2.4 provides evidence that this is not the case.

[22] When looking at the number of votes received as the dependent variables, this last result is only marginally significant.

Result 2 *Participants who belong to the top 25% are more likely to be chosen as leader in the SI treatment than in the AI treatment (supports Hypothesis 2).*

4.3.3 Results on Leader's Decisions

We showed in section 4.3 that leaders in the AI treatment are less likely to belong in the top 25% than leaders in the SI treatment. By design, participants in the AI treatment should end up with lower payoff from their leader's decisions. However, it is still possible for leaders who do not belong to the top 25% to maximize their earning by always choosing the safe option (Option B) over the risky option (Option A). Table 4.6 shows that while 67.74% leaders in the AI treatment do not belong to the top 25%, only 18.52% of them choose Option B from the start.[23]

Table 4.6: Distribution of leaders' switching points.

	switch									
	1	2	3	4	5	6	7	8	9	10
SI	4.17%	0.00%	0.00%	0.00%	20.83%	25.00%	20.83%	20.83%	0.00%	8.33%
AI	18.52%	0.00%	0.00%	7.41%	22.22%	25.93%	14.81%	3.70%	3.70%	3.70%

Note: Table 4.6 shows for each decision, the proportion of leaders who switched from the risky to the safe option at the given decision.

Table 4.7 displays the average switching point from Option A to Option B and the average payoff received from the leaders decisions (with standard errors in parentheses). In both treatments, the average switching point (converted to a scale of 1 to 100) is lower than the average reported beliefs after the chat (65.4% vs 76.7% in the SI treatment and 51.9% vs 67.2% in the AI treatment) and the difference is significant (W test: $p = 0.005$ for both treatments). However, the difference in difference between treatment is not significant (Mann-Whitney test: $p = 0.484$),[24] suggesting that the difference is driven by risk aversion and not by confusion.

In the SI treatment, about 62% of leaders are top-ranked. If leaders were well-calibrated (i.e. leaders' beliefs about their performance are unbiased), the 62% of leaders who are top-ranked should choose Option A from decision 1 to 10, and the remaining 28% should choose Option B from decision 1 to 10. in this case, the average switch would be 6.2. Hence, if leaders are well-calibrated, the average

[23] In our data, 8 leaders exhibit non-monotonic preferences (4 in the SI treatment and 4 in the AI treatment). Results in this section include data for leaders with a unique switching point only. For consistency, we also dropped one leader who obtained one of the best score in the group but was not ranked in the top 25%.

[24] MW test, hereafter.

switch should be equal to the proportion of top-ranked leaders in the treatment divided by 10. If leaders are overconfident on average, they would hold on to Option A longer than they should and the observed switch would be higher than the proportion of top-ranked leaders. In the SI treatment, there is no significant difference between the average switch and the proportion of top-ranked leaders (6.54 vs. 5.86; W test: $p = 0.577$). On the contrary, the average switch of leaders in the AI treatment is significantly different from the proportion of top-ranked leaders (5.19 vs. 3.23; W test: $p = 0.043$), suggesting that leaders in the AI treatment are overconfident on average.

Table 4.7: Summary statistics on leaders' decisions and outcome.

	SI	AI	Δ	p-value
Average switch	6.54	5.19	-1.35	0.030
	(0.381)	(0.468)	(0..613)	
N	24	27	51	
Average payoff from decisions	7.98	5.42	-2.56	< 0.001
	(0.329)	(0.348)	(0.482)	
N	96	108	204	

Note: Table 4.7 displays the leaders' average switching point from the risky option to the safe option and participants' average payoffs from the leaders' decisions. Standard errors in parentheses. We report the p-values of two-sample Mann-Whitney tests between treatments.

If leaders who do not belong to the top 25% were only pretending/bluffing and were not actually believing that they belong to the top 25%, we would have observed (i) a much higher proportion of leaders in the AI treatment choosing the certain amount from the start; and (ii) a much lower average switching point. Thus, results from the AI treatment are consistent with the idea of self-deception. The consequences therefore are lower average payoff from the leaders' decisions in the AI treatment compared to the SI treatment (7.98 AUD vs. 5.42 AUD; two-sample MW test: $p < 0.001$).

In order to investigate the causal treatment effect on the payoff received from the leader, we estimate OLS regressions in which the dependent variable is the payoff received from the leader's decisions. The independent variables include a treatment dummy. Results are reported in column (1) of Table 4.8. In column (2), we control for leaders' demographics (sex, age, risk preferences and social dominance). All regressions are clustered at the group level. Table 4.8 shows that giving a bonus to the leader decreases participants' payoff from their leader's decisions by more than 2 AUD and the results are significant at the 1% level.

Table 4.8: Treatment effect on payoff.

Dep. variable:	Payoff from the leader's decisions	
	(1)	(2)
Treatment	-2.281***	-2.344***
	(0.733)	(0.738)
Leader AMS score		0.113
		(0.083)
Female leader		0.272
		(0.762)
Leader age		0.050
		(0.055)
Leader Risk pref.		-0.234
		(0.221)
Constant	7.483***	4.732*
	(0.562)	(2.588)
N cluster	60	60
N	240	240

Note: Table 4.8 shows the OLS estimates of the treatment effect on the payoff received from the leader. Standard errors in parentheses. *** p <0.01, ** p <0.05, * p <0.10.

To summarize, our results show that leaders in the AI treatment are overconfident: they switch from the risky lottery to the safe outcome later than they should if they were well-calibrated. As a result, participants in the AI treatment earn significantly less on average from their leaders than participants in the SI treatment.

Result 3 *Leaders in the Asymmetric Incentives treatment make overconfident decisions and earn less for their group than leaders in the Symmetric Incentives treatment (supports Hypothesis 3).*

4.4 General Discussion and Conclusion

4.4.1 Discussion

While we observe that participants in both treatments are overconfident regarding their absolute and relative performance, we do not find any significant difference in beliefs between treatments. We explore two explanations for this absence of treatment differences: quartile-specific treatment effect, lack of statistical power and self-image concerns.

First, the absence of treatment effect on beliefs can be due to a difference in the incentives to self-deceive between participants at different point of the distribu-

tions of performances. The AI treatment creates a conflict for non top-ranked participants between being chosen as the leader to earn the additional bonus B and choosing the top-ranked group member as the group leader to maximize $w * score_{1,l} + p_l$. Participant i should engage in self-deceiving strategies if the benefit of doing so outweigh the costs, i.e. if $B > w * (score_{1,l} - score_{1,i}) + p_l - p_i$. Hence, the lower the performance, the more costly it is to self-deceive. If this is the case, we should expect greater overconfidence from participants in the upper quartiles of the distribution of performances, relative to participants in the bottom quartiles. However, analyses of treatment differences in beliefs between the different quartiles do not support this explanation (see Appendix C.2.2).

The modest sample size in our experiment (29 clusters in the SI treatment and 31 clusters in the AI treatment) may have played a role in limiting the significance of some of our results. To check whether our non-significant results were due to a lack of statistical power, we conducted post hoc power analyses using *G*Power* (Faul et al., 2009) with power set at the recommended 0.80 level (Cohen, 2013) and the level of significance at 0.05, two-tailed. This showed us that sample sizes would have to increase up to 506 and 460 clusters in order for differences between treatments for overconfidence and overplacement, respectively, to reach statistical significance, considering leaders only.

Finally, the absence of significant differences in beliefs between both treatments can be driven by the high level of overconfidence observed in the SI treatment. An explanation for this high level of overconfidence in the SI treatment could be that participants form positively biased beliefs about themselves to maintain self-esteem (Schelling, 1988; Bénabou and Tirole, 2002; Köszegi, 2006). In our experiment, participants undertake Raven's standard progressive matrices. This task has been shown to be correlated with IQ (Flynn, 1987; O'Leary et al., 1991; Lynn and Irwing, 2004). Eil and Rao (2011) and Mobius et al. (2011) experimentally show that participants are reluctant to obtain negative information about their intelligence, suggesting that a low performance on this task can be particularly threatening for participants' self-image. If this is the case, participants will form overconfident beliefs even in the absence of direct monetary gains from doing so. Consistent with this idea, Charness et al. (2018) and Schwardman and van der Weele (2019) also observe substantial overconfidence in their treatment in which there is no strategic incentives for participants to over-report their performance at the progressive Raven's matrices.

4.4.2 Conclusion

The literature provide evidence that overconfidence is common among leaders (Malmendier and Tate, 2005; Phua et al., 2018). One reason that could explain this phenomenon is that overconfidence helps individuals to reach higher status (Anderson et al., 2012). In this paper, we investigate whether overconfidence facilitates access to leadership position primarily when leadership provides privileged access to valued resources. To do so, we design an experiment in which we manipulate participants' incentive to become the leader of their group. First, participants undertake an individual task and are ranked according to their performance. In the AI treatment, the leader receives a bonus on top of other payments. In the SI treatment, the payoff function is the same for every group member, no matter their role. This design allow us to test (i) whether overconfidence facilitates promotion to the leadership position; (ii) whether it emerges primarily when being the leader provide access to extra resources and (iii) how overconfidence affects leaders' decisions.

First, we find a substantial level of overconfidence in both treatments. We find that participants who hold higher beliefs also have higher chances to be selected as the leader of their group in both treatments. However, participants in the SI treatment are more likely to choose the top-ranked member in their group as their leader (58.6% of the time) than participants in the AI treatment (32.3% of the time). In addition, we find that in the AI treatment, participants who belong to the second quartile are more likely to be selected as the leader than participants in the first quartile. Regarding the leaders' decisions, leaders in the AI treatment make overconfident decisions on average while leaders in the SI treatment are well-calibrated. More specifically, even though only 32.26% of leaders in AI are the best performer of their group, only 18.52% play the optimal strategy (i.e. choosing the safe option straight from the start) Consequently, participants in the SI treatment earn significantly more than participants in the AI treatment from their leaders' decisions. However, we find little evidence that these differences in leaders' selection between treatments is driven by overconfidence.

Our findings altogether highlight the importance of incentives in the selection of leaders, which has important implications in the real world. Powerful positions often come with high salaries and/or bonuses that are designed to reward extraordinary performances. In our lab experiment, we found that a bonus as small as \$5 is sufficient to make the rate of "good" leader drop by more than 40%. If such rewards are likely to lead to the promotion of less performing individuals, one can wonder whether these advantages are justified.

Chapter 5

General Conclusion

5.1 Summary and Contribution

While a growing body of evidence now support the fact that individuals form positively biased beliefs as a strategy in social interaction, our understanding of this phenomenon remains limited. Using a series of laboratory experiments, this thesis aims (i) to show that overconfidence emerges primarily in strategic interactions when doing so can lead to higher expected payoffs in three different settings and (ii) causally identify the individual and social impact of strategic overconfidence.

In Chapter 2, I experimentally investigates whether overconfidence emerges in social interactions primarily when it provides an advantage. In this experiment, participants are incentivized either to form accurate beliefs about their performance at a test, or to convince a group of other participants that they performed well. I also vary participants' ability to freely gather information about their performance. Results provide, for the first time altogether, the different empirical links of von Hippel and Trivers (2011)'s theory of strategic confidence. Results show that participants who expect to have to convince others about their ability (i) selectively search for information in a way that is conducive to receiving more positive information on their performance; (ii) are more overconfident than those who do not and - in turn - (iii) more successful at convincing others that they performed well.

Building on the findings of Chapter 2, the second part of this thesis investigates situations in which overconfidence can be individually beneficial, but at the expense of the social welfare. In Chapter 3 and 4, I focus on two types of strategic

interactions that highlight this conflict between individual self-interest and social interest: bilateral negotiations and the selection of leaders.

In Chapter 3, I investigate the effect of an exogenous variation in confidence on the outcome of a bilateral bargaining situation. I designed a lab experiment in which I manipulate participants' confidence in a bargaining setting. Participants are matched in pairs and each pair performs a task to earn points on a group account. At the end of the task, I manipulate participants' confidence about their performance relative to their partner's. The group account is then divided in two unequal shares (70/30 percent of the group account) and each pair is asked to agree on how to allocate the shares between its members. I use a 3-stage negotiation process in which participant are given an opportunity to reach an agreement in each of the 3 stages. However, the longer each pair takes to agree, the lower the final amount of each share will be. If the pair fails to reach an agreement, both members end up empty-handed. Results show that an increase in confidence increases the likelihood of impasses and costly delays, which are detrimental for the social outcome of the negotiation. In contrast, I find that participants who are relatively more confident than their partner end up with larger earnings. All together, these findings provide an explanation regarding the persistence of overconfidence in this context.

Finally, Chapter 4 investigates whether individuals become strategically overconfident when being a leader give access to valuable resources and to what extent overconfident leaders are detrimental to the group. I designed a lab experiment in which I exogenously manipulate the incentives to be the leader of a group. After completing an effort task, participants are matched in groups of four. Each group have to select a leader that will make a series of risky binary choices on behalf of the group. For each decision, the probability to obtain the desirable outcome depends on the likelihood that the leader was the best performer in his group. Thus, selecting the best performer of the group as the leader maximizes the social outcome. Before making their choice of leader, group members are allowed to communicate among them via a group chat. In one treatment, the payoff function of the leader is the same as the other group members. In another treatment, the leader receives a bonus on top of his payoff while the payoff function of the other group members remains the same. Results show that leaders in the treatment with the bonus (i) are less likely to be the best performer in their group and (ii) make overconfident choices that earn less money for their group compared to leaders who did not receive the bonus.

The contributions of this thesis are twofold. First, I provide compelling evidence

that overconfidence emerges in social interactions primarily when it provides an advantage. This finding enhances our understanding of the situational determinants of overconfidence in social interactions: overconfidence is more likely to arise in competitive situations in which there is room for deception and true ability is not observable or its observation is noisy. This finding has important implications for economic theory, especially the handling of beliefs in theoretical models. Most standard theories assume that individuals are Bayesian-rational, meaning that people collect and process information in a way that gives them a relatively accurate perception of reality. If, instead, beliefs are "chosen", this assumption is clearly violated. This thesis helps to identify when these assumptions may not be warranted, such as situations in which the strategic value of overconfidence is high.

Second, I provide an explanation for the persistence of overconfidence in such situations despite its costs. The existence of strategic benefits generates a demand for inflated beliefs that leads to inefficiencies when individuals' private interest conflicts with the social interest. Although this thesis sits squarely in the domain of basic research, it addresses real-world issues that are particularly relevant to this day. In the past couple years, overconfidence has been blamed for major world-changing events such as Brexit or the election of President Donald Trump, described by Walter Shapiro as "*the most laughable manifestation of overconfidence in the 2018 campaign*".[1] These events already have and will have economic and financial consequences all around the world and stressed the importance of understanding overconfidence to better prevent its rising. The findings from this thesis have the potential to help organizations and policy makers identify situations in which overconfidence is likely to be costly for society; and lay the foundations to improve policies intended to prevent or limit its negative effects.

5.2 Shortcomings and Extensions

5.2.1 The Trouble with Overconfidence

Early experimental results in psychology document a substantial level of overconfidence in a wide range of domains. However, subjects claims were self-reported and not verifiable by the experimenter. For instance, in Svenson (1981)'s experiment, participants were asked to report how well they think they drive compared to the average driver while the experimenters had no way to measure the actual driving skills of their subjects. A growing body of experiments now measure over-

[1]The article can be found online here: https://www.rollcall.com/news/opinion/republicans-fell-trumps-confidence-game.

confidence in a way that is (i) incentive-compatible and (ii) quantifiable in the lab. In these experiments, participants perform an effort task. Overconfidence is then mainly measured in two ways: (i) overestimation of one's actual performance, (ii) overplacement of one's performance relative to others. (i) is usually measured as the discrepancy between participants' performance at the task and their beliefs about their performance. Economists favors the Stochastic Becker-DeGroot-Marschak belief elicitation mechanism (SBDM) (Becker et al., 1964) to measure (ii) because it relaxes the assumption of risk neutrality. Participants make a series of binary choices. Participants can choose to be paid according to the realization of a probabilistic event A. The participant receives a prize k if the event is realized, and 0 otherwise. For example, A can be "participant i's performance belong to the top 50% of performance in the session". Or participants can choose a lottery ticket that gives them X% chance to receive the prize k. Participants are asked to state for what value of X they would be indifferent between being paid according to the first option and playing the lottery. Thus, X indicate participants' beliefs. While both methods provide robust findings of the persistence of overconfidence, they have their own limitations.

Participants usually undertake a known number of tasks n. Since n is finite, participants' performance is bounded. Thus, high ability individuals have less room (or no room at all) to form positively biased beliefs about their performance than medium and low ability individuals. In Chapter 3, we try to overcome this issue by giving participants a time to complete as many tasks as possible, eliminating the upper bound of possible performance. However, doing so moves the boundary problem to the left hand side of the distribution as it is know easier for participants with low performance to estimate their performance and consequently, harder for them to self-deceive. SBDM mechanisms do not only suffer from the same issue, they are also non trivial. They are now evidence that participants need a proper training to understand how these mechanisms work (Burfurd and Wilkening, 2018).

Establishing causal relationships between beliefs and behavior also requires to exogenously manipulate those beliefs in the laboratory. However, if performance - and consequently beliefs - is bounded, the extent to which beliefs can be manipulated is ultimately bounded too. Examples of experiments that provide such manipulations in economics are scarce (Schwardman and van der Weele, 2019; Barron and Gravert, 2018; Charness et al., 2019). From this small literature, we can identify two main methods. The first one is to manipulate the difficulty of the task (easy vs. hard). Based on evidence that participants are usually overconfident regarding their relative performance on easy task and underconfident

on difficult task (Moore and Healy, 2008; Larrick and Soll, 2007), beliefs of participants who undertake the easy version of the task should be higher on average than beliefs of participants who undertake the difficult version of the task. The second one is to provide participants with a noisy binary signals about their relative performance.

In Chapter 4, I attempted to manipulate participants' beliefs by manipulating the monetary incentives of being overconfident. However, the effect of these manipulations is often limited. First, participants may have strong priors about their abilities, especially if the tasks at hand is self-relevant. For example, experiments using progressive Raven's matrices that are well-known to correlates with IQ find a substantial level of overconfidence in the control group (Charness et al., 2018; Schwardman and van der Weele, 2019) which in turn makes it harder to shift the distribution of beliefs in the treated group.

In summary, the proper elicitation and manipulation of beliefs depend heavily on the task chosen by the experimenters. The absence of a defined framework to study overconfidence makes between-study comparisons difficult and challenge the external validity of the experimental findings. Developing a unique methodology to elicit and manipulate beliefs in the laboratory seems an important next step for future research investigating overconfidence. Based on the limitations identified from the small sample of previous studies, this ideal framework would include (i) a task that allows both extreme of the distribution to self-deceive and (ii) an elicitation mechanism that can be easily understood by participants.

5.2.2 Moral vs. Immoral Leadership

In Chapter 4, we attempt to shift upwards participants' beliefs about their performance by providing an extra bonus for being chosen as the group leader. A by product of this manipulation is that in creates vertical inequality of resource distribution. Thus, Chapter 4 also suggests that income inequality favors the emergence of "immoral" leadership (i.e. the act of making decisions that benefit the leader at the expense of the group). However, in most modern organizations the leader gains benefits beyond the rank-and-file members. Thus, the pragmatic importance of inequality is not clear from this experiment, as almost all modern organizations and political entities would be classified as unequal and virtually none would be considered equal by such an approach. Hence, I propose a follow up lab experiment that will allow to investigate whether more unequal organizations facilitate (i) immoral leadership and (ii) poor leadership selection, and (iii) the role of overconfidence in the leader selection process.

As in the experiment proposed in Chapter 4, participants will perform an effort task and be matched in groups of four depending on their performance at the task. However, rather than choosing a single leader, participants will have to choose management roles for each member of their role. I then manipulate the level of inequality between the different levels of management within each group. In the "low inequality" treatment, each increase in level will be associated with a 10% increase in the size of the bonus payment. In the "high inequality" treatment, each increase in level leads to a 50% increase in the size of the bonus. Besides the bonus, the payoff function remains the same as in Chapter 4 to ensure that it is socially optimal for the group to select higher performing members in higher positions. Once the leadership structure has been chosen, participants on the management team will be confronted with the same set of choices as in Chapter 4. and will meet as a group to decide which decisions to make. This design allows to test how closely the final ranking of the team members matches their performance in the first phase of the experiment, as well as the degree to which the leadership team makes well-calibrated or overconfident decisions that do or do not maximize group outcomes (as in Chapter 4). In the low inequality treatment, self-interest is yoked to group-interest: any benefits that a skilled and capable leader might bring are group-level benefits. Moreover, the leadership role garner fewer personal privileges as there is relatively less to gain (and lose). Thus, this type of organization should facilitate the selection of moral leaders compared to a more unequal environment.

Leadership quality is one of the most important determinants of organizational outcomes (Haslam et al., 2010). Immoral and self-serving leaders can quickly transform highly functional and effective organizations into dysfunctional, corrupt, and inefficient entities (Probst and Raisch, 2005). Jong-Sung and Khagram (2005) conducted a large-scale comparative analysis of 129 countries to probe the predictors of corruption and found the positive relationship between inequality and corruption to be at least as important as the role of economic development and natural resource abundance. By demonstrating when and how inequality leads to immoral leadership, this follow up experiment will play an important role in demonstrating the preconditions that enable effective and cooperative organizations and societies.

Bibliography

Abeler, J., Becker, A., and Falk, A. (2014). Representative evidence on lying costs. *Journal of Public Economics*, 113:96–104.

Abeler, J., Nosenzo, D., and Raymond, C. (2016). Preferences for truth-telling.

Adams, J. S. (1965). Inequity in social exchange. In *Advances in experimental social psychology*, volume 2, pages 267–299. Elsevier.

Akerlof, G. A. (1970). The market for" lemons": Quality uncertainty and the market mechanism, 84q. *J. ECON*, 488:489–90.

Alaoui, L. (2009). The value of useless information.

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of personality and social psychology*, 49(6):1621.

Alloy, L. B. and Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of experimental psychology: General*, 108(4):441.

Ambuehl, S. and Li, S. (2018). Belief updating and the demand for information. *Games and Economic Behavior*, 109:21–39.

Anderson, C., Brion, S., Moore, D. A., and Kennedy, J. A. (2012). A status-enhancement account of overconfidence. *Journal of Personality and Social Psychology*, 103(4):718–735.

Andolfatto, D., Mongrain, S., and Myers, G. M. (2005). Self-esteem and labour market choices.

Arbak, E. and Villeval, M.-C. (2013). Voluntary leadership: Motivation and influence. *Social Choice and Welfare*, 40(3):635–662.

Arechar, A. A., Gächter, S., and Molleman, L. (2018). Conducting interactive experiments online. *Experimental economics*, 21(1):99–131.

Babcock, L., Loewenstein, G., Issacharoff, S., and Camerer, C. (1995a). Biased judgments of fairness in bargaining. *The American Economic Review*, 85(5):1337–1343.

Babcock, L., Loewenstein, G., and Wang, X. (1995b). The relationship between uncertainty, the contract zone, and efficiency in a bargaining experiment. *Journal of Economic Behavior & Organization*, 27(3):475–485.

Bar-Gill, O. (2005). The evolution and persistence of optimism in litigation. *Journal of Law, Economics, and Organization*, 22(2):490–507.

Barber, B. and Odean, T. (2001). Boys will be boys: Gender, overconfidence and common stock investment. *The Quarterly Journal of Economics*, 116(1):261–292.

Barron, K. and Gravert, C. (2018). Confidence and Career Choices: An Experiment. *SSRN Electronic Journal*.

Baumeister, R. F., Tice, D. M., and Hutton, D. G. (1989). Self-presentational motivations and personality differences in self-esteem. *Journal of personality*, 57(3):547–579.

Bazerman, M. H. and Neale, M. A. (1982). Improving negotiation effectiveness under final offer arbitration: The role of selection and training. *Journal of Applied Psychology*, 67(5):543.

Becker, G. M., DeGroot, M. H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral science*, 9(3):226–232.

Becker, G. S. (1968). Crime and punishment: An economic approach. In *The economic dimensions of crime*, pages 13–68. Springer.

Belot, M. and Van de Ven, J. (2017). How private is private information? the ability to spot deception in an economic game. *Experimental economics*, 20(1):19–43.

Bénabou, R. (2012). Groupthink: Collective delusions in organizations and markets. *Review of Economic Studies*, 80(2):429–462.

Bénabou, R. (2015). The Economics of Motivated Beliefs. *Revue d'économie politique*, 125(5):665–685.

Bénabou, R. and Tirole, J. (2002). Self-confidence and personal motivation. *The Quarterly Journal of Economics*, 117(3):871–915.

Bénabou, R. and Tirole, J. (2004). Willpower and personal rules. *Journal of Political Economy*, 112(4):848–886.

Bénabou, R. and Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics*, 126(2):805–855.

Bénabou, R. and Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3):141–64.

Benos, A. V. (1998). Aggressiveness and survival of overconfident traders. *Journal of Financial Markets*, 1(3-4):353–383.

Bentham, J. (1789). Introduction to the principles of morals and legislation. 1996 imprint.

Berger, J., Cohen, B., and Zelditch, M. (1972). Status characteristics and social interaction. *American Sociological Review*, 37:241–255.

Berglas, S. and Jones, E. E. (1978). Drug choice as a self-handicapping strategy in response to noncontingent success. *Journal of personality and social psychology*, 36(4):405.

Bing, M. N. and Davidson, H. (2012). Measuring faking using the overclaiming instrument. River Cities Industrial and Organizational Psychology Conference, Chattanooga, TN.

Blaker, N. M., Rompa, I., Dessing, I. H., Vriend, A. F., Herschberg, C., and van Vugt, M. (2013). The height leadership advantage in men and women: Testing evolutionary psychology predictions about the perceptions of tall leaders. *Group Processes and Intergroup Relations*, 16(1):17–27.

Blau, P. (1964). *Power and Exchange in Social Life*. Ny: John wiley & sons edition.

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.

Boles, T. L., Croson, R. T., and Murnighan, J. K. (2000). Deception and retribution in repeated ultimatum bargaining. *Organizational behavior and human decision processes*, 83(2):235–259.

Bond, C. F. and DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and social psychology Review*, 10(3):214–234.

Bond, C. F. and DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological bulletin*, 134(4):477.

Boulu-Reshef, B., Holt, C. A., and Thomas-Hunt, M. C. (2015). Organization Style, Leadership Strategy and Free-Riding. *SSRN Electronic Journal*.

Brandts, J. and Cooper, D. J. (2007). It's What You Say, Not What You Pay: An Experimental Study of Manager-Employee Relationships in Overcoming Coordination Failure. *Journal of the European Economic Association*, 5(6):1223– 1268.

Brandts, J., Cooper, D. J., and Fatas, E. (2007). Leadership and overcoming coordination failure with asymmetric costs. *Experimental Economics*, 10(3):269–284.

Brandts, J., Cooper, D. J., and Weber, R. A. (2015). Legitimacy, Communication, and Leadership in the Turnaround Game. *Management Science*, 61(11):2627–2645.

Brandts, J., Rott, C., and Solà, C. (2016). Not just like starting over - Leadership and revivification of cooperation in groups. *Experimental Economics*, 19(4):792–818.

Brock, T. C. and Balloun, J. L. (1967). Behavioral receptivity to dissonant information. *Journal of personality and social psychology*, 6(4p1):413.

Buller, D. B. and Burgoon, J. K. (1996). Interpersonal deception theory. *Communication theory*, 6(3):203–242.

Burfurd, I. and Wilkening, T. (2018). Experimental guidance for eliciting beliefs with the stochastic becker–degroot–marschak mechanism. *Journal of the Economic Science Association*, 4(1):15–28.

Burks, S. V., Carpenter, J. P., Goette, L., and Rustichini, A. (2010). Overconfidence is a social signaling bias.

Buser, T., Gerhards, L., and Van der Weele, J. J. (2016). Measuring responsiveness to feedback as a personal trait.

Buss, D. M. (2009). The great struggles of life: Darwin and the emergence of evolutionary psychology. *American Psychologist*, 64(2):140.

Bénabou, R. and Tirole, J. (2016). Mindful Economics: The Production, Consumption, and Value of Beliefs. *Journal of Economic Perspectives*, 30:141–164.

Camerer, C. and Lovallo, D. (1999). Overconfidence and excess entry: An experimental approach. *American economic review*, 89(1):306–318.

Cantor, N. and Norem, J. K. (1989). Defensive pessimism and stress and coping. *Social cognition*, 7(2):92–112.

Carver, C. S. and Scheier, M. F. (2014). Dispositional optimism. *Trends in cognitive sciences*, 18(6):293–299.

Case, K. E. and Shiller, R. J. (2003). Is there a bubble in the housing market? *Brookings papers on economic activity*, 2003(2):299–342.

Cassidy, T. and Lynn, R. (1989). A multifactorial approach to achievement motivation: The development of a comprehensive measure. *Journal of Occupational Psychology*.

Chambers, D. and Reisberg, D. (1985). Can mental images be ambiguous? *Journal of Experimental Psychology: Human perception and performance*, 11(3):317.

Chambers, J. R. and Windschitl, P. D. (2004). Biases in social comparative judgments: the role of nonmotivated factors in above-average and comparative-optimism effects. *Psychological bulletin*, 130(5):813.

Charness, G. and Dave, C. (2017). Confirmation bias with motivated beliefs. *Games and Economic Behavior*, 104:1–23.

Charness, G., Naef, M., and Sontuoso, A. (2019). Opportunistic conformism. *Journal of Economic Theory*, 180:100–134.

Charness, G., Rustichini, A., and Van de Ven, J. (2018). Self-confidence and strategic behavior. *Experimental Economics*, 21(1):72–98.

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Chew, S. H., Huang, W., and Zhao, X. (2018). Motivated false memory. *Mimeo, National University of Singapore, February*.

Christ, S. E., Van Essen, D. C., Watson, J. M., Brubaker, L. E., and McDermott, K. B. (2008). The contributions of prefrontal cortex and executive control to deception: evidence from activation likelihood estimate meta-analyses. *Cerebral cortex*, 19(7):1557–1566.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Cohn, A., Fehr, E., and Maréchal, M. A. (2014). Business culture and dishonesty in the banking industry. *Nature*, 516(7529):86.

Colwell, K., Hiscock-Anisman, C., Memon, A., Colwell, L. H., Taylor, L., and Woods, D. (2009). Training in assessment criteria indicative of deception to improve credibility judgments. *Journal of Forensic Psychology Practice*, 9(3):199–207.

Conway, M. and Ross, M. (1984). Getting what you want by revising what you had. *Journal of personality and social psychology*, 47(4):738.

Coutts, A. (2019). Good news and bad news are still news: Experimental evidence on belief updating. *Experimental Economics*, 22(2):369–395.

Crary, W. G. (1966). Reactions to incongruent self-experiences. *Journal of Consulting Psychology*, 30(3):246.

Cross, K. P. (1977). Not can, but will college teaching be improved? *New Directions for Higher Education*, 1977(17):1–15.

Croyle, R. T., Loftus, E. F., Barger, S. D., Sun, Y.-C., Hart, M., and Gettig, J. (2006). How well do people recall risk factor test results? accuracy and bias among cholesterol screening participants. *Health Psychology*, 25(3):425.

Dawson, E., Gilovich, T., and Regan, D. T. (2002). Motivated reasoning and performance on the was on selection task. *Personality and Social Psychology Bulletin*, 28(10):1379–1387.

Dawson, E., Savitsky, K., and Dunning, D. (2006). "don't tell me, i don't want to know": Understanding people's reluctance to obtain medical diagnostic information 1. *Journal of Applied Social Psychology*, 36(3):751–768.

De Jong, A., De Ruyter, K., and Wetzels, M. (2006). Linking employee confidence to performance: A study of self-managing service teams. *Journal of the Academy of Marketing Science*, 34(4):576–587.

DePaulo, B. M. (2004). The many faces of lies. *The social psychology of good and evil, ed. AG Miller*, pages 303–26.

DePaulo, B. M. and Kashy, D. A. (1998). Everyday lies in close and casual relationships. *Journal of personality and social psychology*, 74(1):63.

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., and Epstein, J. A. (1996). Lying in everyday life. *Journal of personality and social psychology*, 70(5):979.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., and Cooper, H. (2003). Cues to deception. *Psychological bulletin*, 129(1):74.

DePaulo, B. M., Stone, J., and Lassiter, D. (1985). *Deceiving and detecting deceit. In B. Schenkler (Ed.)*. The self and social life.

Descamps, A., Massoni, S., and Page, L. (2016). Knowing when to stop and make a choice, an experiment on optimal sequential sampling.

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.

Dreber, A. and Johannesson, M. (2008). Gender differences in deception. *Economics Letters*, 99(1):197–199.

Drouvelis, M. and Nosenzo, D. (2013). Group identity and leading-by-example. *Journal of Economic Psychology*, 39:414–425.

Dunning, D., Heath, C., and Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological science in the public interest*, 5(3):69–106.

Dunning, D., Leuenberger, A., and Sherman, D. A. (1995). A new look at motivated inference: Are self-serving theories of success a product of motivational forces? *Journal of Personality and Social Psychology*, 69(1):58.

Eil, D. and Rao, J. M. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2):114–38.

Ekman, P., Friesen, W. V., and O'sullivan, M. (1988). Smiles when lying. *Journal of personality and social psychology*, 54(3):414.

Ekman, P. and O'sullivan, M. (1991). Who can catch a liar? *American psychologist*, 46(9):913.

Eliaz, K. and Schotter, A. (2010). Paying for confidence: An experimental study of the demand for non-instrumental information. *Games and Economic Behavior*, 70(2):304–324.

Ellis, L. E. (1994). *Social stratification and socioeconomic inequality, Vol. 2: Reproductive and interpersonal aspects of dominance and status.* Praeger publishers/greenwood publishing group edition.

Epley, N. and Whitchurch, E. (2008). Mirror, mirror on the wall: Enhancement in self-recognition. *Personality and Social Psychology Bulletin*, 34(9):1159–1170.

Erat, S. and Gneezy, U. (2012). White lies. *Management Science*, 58(4):723–733.

Ertac, S. (2011). Does self-relevance affect information processing? experimental evidence on the response to performance and non-performance feedback. *Journal of Economic Behavior & Organization*, 80(3):532–545.

Evans III, J. H., Hannan, R. L., Krishnan, R., and Moser, D. V. (2001). Honesty in managerial reporting. *The Accounting Review*, 76(4):537–559.

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.

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.

Flynn, J. R. (1987). Massive iq gains in 14 nations: What iq tests really measure. *Psychological bulletin*, 101(2):171.

Frackenpohl, G., Hillenbrand, A., and Kube, S. (2016). Leadership effectiveness and institutional frames. *Experimental Economics*, 19(4):842–863.

Frey, D. (1986). Recent research on selective exposure to information. In *Advances in experimental social psychology*, volume 19, pages 41–80. Elsevier.

Gabriel, M., Critelli, J., and Ee, J. (1994). Narcissistic illusions in self-evaluations of intelligence and attractiveness. *Journal of Personality*, 62(1):143–155.

Gächter, S. and Riedl, A. (2005). Moral property rights in bargaining with infeasible claims. *Management Science*, 51(2):249–263.

Galasso, A. and Simcoe, T. S. (2011). Ceo overconfidence and innovation. *Management Science*, 57(8):1469–1484.

Galinsky, A. D. and Mussweiler, T. (2001). First offers as anchors: the role of perspective-taking and negotiator focus. *Journal of personality and social psychology*, 81(4):657.

Ganguly, A. and Tasoff, J. (2016). Fantasy and dread: The demand for information and the consumption utility of the future. *Management Science*, 63(12):4037–4060.

Gillet, J., Cartwright, E., and van Vugt, M. (2011). Selfish or servant leadership? Evolutionary predictions on leadership personalities in coordination games. *Personality and Individual Differences*, 51(3):231–236.

Ging-Jehli, N. R., Schneider, F., and Weber, R. A. (2019). On self-serving strategic beliefs. *University of Zurich, Department of Economics, Working Paper*, (315).

Glaeser, E. L. and Sunstein, C. R. (2013). Why does balanced news produce unbalanced views? Technical report, National Bureau of Economic Research.

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

Gneezy, U., Kajackaite, A., and Sobel, J. (2018). Lying aversion and the size of the lie. *American Economic Review*, 108(2):419–53.

Goel, A. M. and Thakor, A. V. (2008). Overconfidence, ceo selection, and corporate governance. *The Journal of Finance*, 63(6):2737–2784.

Golman, R., Hagmann, D., and Loewenstein, G. (2017). Information avoidance. *Journal of Economic Literature*, 55(1):96–135.

Gottlieb, D. (2014). Imperfect memory and choice under risk. *Games and Economic Behavior*, 85:127–158.

Griskevicius, V., Tybur, J. M., and Van den Bergh, B. (2010). Going green to be seen: Status, reputation, and conspicuous conservation. *Journal of Personality and Social Psychology*, 98(3):392–404.

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.

Gur, R. C. and Sackeim, H. A. (1979). Self-deception: A concept in search of a phenomenon. *Journal of Personality and Social Psychology*, 37(2):147.

Gächter, S., Nosenzo, D., Renner, E., and Sefton, M. (2012). Who makes a good leader? Cooperativeness, optimism and leading-by-example. *Economic Inquiry*, 50(4):953–967.

Gächter, S. and Renner, E. (2014). Leaders as Role Models for the Voluntary Provision of Public Goods. *CESifo WP No. 5049*.

Güth, W., Levati, M. V., Sutter, M., and van der Heijden, E. (2007). Leading by example with and without exclusion power in voluntary contribution experiments. *Journal of Public Economics*, 91(5-6):1023–1042.

Haigner, S. D. and Wakolbinger, F. (2010). To lead or not to lead. *Economics Letters*, 108(1):93–95.

Hamman, J. R., Weber, R. A., and Woon, J. (2011). An Experimental Investigation of Electoral Delegation and the Provision of Public Goods: Electoral delegation and public goods. *American Journal of Political Science*, 55(4):738–752.

Hannan, R. L., Rankin, F. W., and Towry, K. L. (2006). The effect of information systems on honesty in managerial reporting: A behavioral perspective. *Contemporary Accounting Research*, 23(4):885–918.

Hartwig, M., Granhag, P. A., Strömwall, L. A., and Vrij, A. (2004). Police officers' lie detection accuracy: Interrogating freely versus observing video. *Police Quarterly*, 7(4):429–456.

Haslam, S. A., Reicher, S. D., and Platow, M. J. (2010). *The new psychology of leadership: Identity, influence and power*. Psychology Press.

Heaton, J. B. (2002). Managerial optimism and corporate finance. *Financial management*, pages 33–45.

Heger, S. A. and Papageorge, N. W. (2013). We should totally open a restaurant: performance uncertainty and optimistic beliefs. Technical report, Mimeo.

Heifetz, A. and Segev, E. (2004). The evolutionary role of toughness in bargaining. *Games and Economic Behavior*, 49(1):117–134.

Heifetz, A., Shannon, C., and Spiegel, Y. (2007). What to maximize if you must. *Journal of Economic Theory*, 133(1):31–57.

Hernandez, R., Kershaw, K. N., Siddique, J., Boehm, J. K., Kubzansky, L. D., Diez-Roux, A., Ning, H., and Lloyd-Jones, D. M. (2015). Optimism and cardiovascular health: multi-ethnic study of atherosclerosis (mesa). *Health behavior and policy review*, 2(1):62–73.

Hirshleifer, D., Low, A., and Teoh, S. H. (2012). Are overconfident ceos better innovators? *The journal of finance*, 67(4):1457–1498.

Hoffman, M. (2016). How is information valued? evidence from framed field experiments. *The Economic Journal*, 126(595):1884–1911.

Hogan, R., Curphy, G. J., and Hogan, J. (1994). What we know about leadership: Effectiveness and personality. *American psychologist*, 49(6):493.

Hong, F., Huang, W., and Zhao, X. (2018). Sunk Cost as a Self-Management Device. *Management Science*.

Hoyle, R. H., Kernis, M. H., Leary, M. R., and Baldwin, M. W. (1999). *Selfhood: Identity, esteem, regulation.* Westview Press.

Huck, S., Szech, N., and Wenner, L. M. (2015). More effort with less pay: On information avoidance, belief design and performance.

Hyman, R. (1989). The psychology of deception. *Annual review of psychology*, 40(1):133–154.

Isaacowitz, D. M. (2006). Motivated gaze: The view from the gazer. *Current Directions in Psychological Science*, 15(2):68–72.

Isaacowitz, D. M., Toner, K., Goren, D., and Wilson, H. R. (2008). Looking while unhappy: Mood-congruent gaze in young adults, positive gaze in older adults. *Psychological Science*, 19(9):848–853.

Ishida, J. (2012). Contracting with self-esteem concerns. *Journal of Economic Behavior and Organization*, 81(2):329–340.

Jack, B. K. and Recalde, M. P. (2015). Leadership and the voluntary provision of public goods: Field evidence from Bolivia. *Journal of Public Economics*, 122:80–93.

Johnson, D. D. (2004). Overconfidence and war: The havoc and glory of positive illusions. Cambridge, MA: Harvard University Press.

Johnson, D. D. and Fowler, J. H. (2011). The evolution of overconfidence. *Nature*, 477(7364):317.

Jones, M. and Sugden, R. (2001). Positive confirmation bias in the acquisition of information. *Theory and Decision*, 50(1):59–99.

Jong-Sung, Y. and Khagram, S. (2005). A comparative study of inequality and corruption. *American sociological review*, 70(1):136–157.

Karlsson, N., Loewenstein, G., and Seppi, D. (2009). The ostrich effect: Selective attention to information. *Journal of Risk and uncertainty*, 38(2):95–115.

Keltner, D., Gruenfeld, D. H., and Anderson, C. (2003). Power, Approach, and Inhibition. *Psychological review*, 110(2):265.

Klofstad, C. A., Anderson, R. C., and Peters, S. (2012). Sounds like a winner: Voice pitch influences perception of leadership capacity in both men and women. *Proceedings of the Royal Society B: Biological Sciences*, 279(1738):2698–2704.

Konow, J. (2003). Which is the fairest one of all? a positive analysis of justice theories. *Journal of economic literature*, 41(4):1188–1239.

Korn, L., Gonen, E., Shaked, Y., and Golan, M. (2013). Health perceptions, self and body image, physical activity and nutrition among undergraduate students in israel. *PloS one*, 8(3):e58543.

Korner, I. N. (1950). *Experimental Investigation of some aspects of the problem of repression: repressive forgetting.* Number 970. Bureau of Publications, Teachers College, Columbia University.

Köszegi, B. (2006). Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, 4(4):673–707.

Kouchaki, M. and Gino, F. (2016). Memories of unethical actions become obfuscated over time. *Proceedings of the National Academy of Sciences*, 113(22):6166–6171.

Kramer, R., Newton, E., and Pommerenke, P. (1993). Self-enhancement biases and negotiator judgment: Effects of self-esteem and mood. *Organizational Behavior and Human Decision Processes*, 56(1):110–113.

Kuhnen, C. M. (2015). Asymmetric learning from financial information. *The Journal of Finance*, 70(5):2029–2062.

Kyle, A. S. and Wang, F. A. (1997). Speculation Duopoly with Agreement to Disagree: Can Overconfidence Survive the Market Test? *The Journal of Finance*, 52(5):2073–2090.

Körding, K. P. and Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244–247.

Larrick, R. P., B. K. A. and Soll, J. (2007). Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not). *Organizational Behavior and Human Decision Processes*, 102(1):76–94.

Lerman, C., Croyle, R. T., Tercyak, K. P., and Hamann, H. (2002). Genetic testing: psychological aspects and implications. *Journal of consulting and clinical psychology*, 70(3):784.

Lerman, C., Hughes, C., Trock, B. J., Myers, R. E., Main, D., Bonney, A., Abbaszadegan, M. R., Harty, A. E., Franklin, B. A., Lynch, J. F., et al. (1999). Genetic testing in families with hereditary nonpolyposis colon cancer. *Jama*, 281(17):1618–1622.

Lerman, C., Narod, S., Schulman, K., Hughes, C., Gomez-Caminero, A., Bonney, G., Gold, K., Trock, B., Main, D., Lynch, J., et al. (1996). Brca1 testing in families with hereditary breast-ovarian cancer: a prospective study of patient decision making and outcomes. *Jama*, 275(24):1885–1892.

Levati, M. V., Sutter, M., and van der Heijden, E. (2007). Leading by Example in a Public Goods Experiment with Heterogeneity and Incomplete Information. *Journal of Conflict Resolution*, 51(5):793–818.

Levy, D. M., Padgitt, K., Peart, S. J., Houser, D., and Xiao, E. (2011). Leadership, cheap talk and really cheap talk. *Journal of Economic Behavior & Organization*, 77(1):40–52.

Lewicki, R. J. (1983). Lying and deception: A behavioral model. *Negotiating in organizations*, 68:90.

Li, K. K. (2013). Asymmetric memory recall of positive and negative events in social interactions. *Experimental Economics*, 16(3):248–262.

Li, K. K. (2017). What determines overconfidence and memory recall bias? the role of feedback, awareness and social comparison. Technical report, Mimeo, University of Hong-Kong.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology.*

Loewenstein, G., Seppi, D., Sicherman, N., and Utkus, S. (2016). Financial attention. *Review of Financial Studies*, 29(4):863–897.

López-Pérez, R. and Spiegelman, E. (2013). Why do people tell the truth? experimental evidence for pure lie aversion. *Experimental Economics*, 16(3):233–247.

Lord, C. G., Ross, L., and Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37(11):2098.

Lundquist, T., Ellingsen, T., Gribbe, E., and Johannesson, M. (2009). The aversion to lying. *Journal of Economic Behavior & Organization*, 70(1-2):81–92.

Lynn, R. and Irwing, P. (2004). Sex differences on the progressive matrices: A meta-analysis. *Intelligence*, 32(5):481–498.

Lyter, D. W., Valdiserri, R. O., Kingsley, L. A., Amoroso, W. P., and Rinaldo Jr, C. R. (1987). The hiv antibody test: why gay and bisexual men want or do not want to know their results. *Public Health Reports*, 102(5):468.

Malmendier, U. and Tate, G. (2005). Behavioral ceos: The role of managerial overconfidence. *The Journal of Economic Perspectives*, 29(4):37–60.

Mann, S. and Vrij, A. (2006). Police officers' judgements of veracity, tenseness, cognitive load and attempted behavioural control in real-life police interviews. *Psychology, Crime & Law*, 12(3):307–319.

Markussen, T. and Tyran, J.-R. (2017). Choosing a public-spirited leader: An experimental investigation of political selection. *Journal of Economic Behavior & Organization*, 144:204–218.

Maslow, A. H. (1950). Self-actualizing people: a study of psychological health. *Personality*.

Mayraz, G. (2011). Wishful thinking. *Available at SSRN 1955644*.

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.

McGillicuddy, N., Pruitt, D., and Syna, H. (1984). Perceptions of firmness and strength in negotiation. *Personality and Social Psychology Bulletin*, 10(3):402–409.

Mischel, W., Ebbesen, E. B., and Zeiss, A. M. (1976). Determinants of selective memory about the self. *Journal of Consulting and Clinical Psychology*, 44(1):92.

Mobius, M. M., Niederle, M., Niehaus, P., and Rosenblat, T. S. (2011). Managing self-confidence. Working paper.

Mobius, M. M., Niederle, M., Niehaus, P., and Rosenblat, T. S. (2014). Managing self-confidence. Working paper.

Mobius, M. M. and Rosenblat, T. S. (2006). Why beauty matters. *American Economic Review*, 96(1):222–235.

Moore, C. (2016). Always the hero to ourselves: The role of self-deception in unethical behavior.

Moore, D. A. (2004). Myopic prediction, self-destructive secrecy, and the unexpected benefits of revealing final deadlines in negotiation. *Organizational Behavior and Human Decision Processes*, 94(2):125–139.

Moore, D. A. and Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, 115(2):502.

Murphy, S. C., Barlow, F. K., and von Hippel, W. (2017). A longitudinal test of three theories of overconfidence. *Social Psychological and Personality Science*, pages 1948–5506.

Murphy, S. C., von Hippel, W., Dubbs, S. L., Angilletta, M. J. J., Wilson, R. S., Trivers, R., and Barlow, F. K. (2015). The role of overconfidence in romantic desirability and competition. *Personality and Social Psychology Bulletin*, 41:1036–1052.

Neale, M. A. and Bazerman, M. H. (1983). The Role of Perspective-Taking Ability in Negotiating under Different Forms of Arbitration. *Industrial and Labor Relations Review*, 36(3):378.

Neale, M. A. and Bazerman, M. H. (1985). The effects of framing and negotiator overconfidence on bargaining behaviors and outcomes. *Academy of Management Journal*, 28(1):34–49.

Novotny, P., Colligan, R. C., Szydlo, D. W., Clark, M. M., Rausch, S., Wampfler, J., Sloan, J. A., and Yang, P. (2010). A pessimistic explanatory style is prognostic for poor lung cancer survival. *Journal of Thoracic Oncology*, 5(3):326–332.

O'Leary, U.-M., Rusch, K. M., and Guastello, S. J. (1991). Estimating age-stratified wais-r iqs from scores on the raven's standard progressive matrices. *Journal of Clinical Psychology*, 47(2):277–284.

Olson, J. M. and Zanna, M. P. (1979). A new look at selective exposure. *Journal of Experimental Social Psychology*, 15(1):1–15.

Ortner, J. (2013). Optimism, delay and (in) efficiency in a stochastic model of bargaining. *Games and Economic Behavior*, 77(1):352–366.

Oster, E., Shoulson, I., and Dorsey, E. (2013). Optimal expectations and limited medical testing: evidence from huntington disease. *American Economic Review*, 103(2):804–30.

Paulhus, D. L., Harms, P. D., Bruce, M. N., and Lysy, D. C. (2003). The over-claiming technique: measuring self-enhancement independent of ability. *Journal of personality and social psychology*, 84(4):890.

Penrod, S. and Cutler, B. (1995). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy, and Law*, 1(4):817.

Phua, K., Tham, T. M., and Wei, C. (2018). Are overconfident ceos better leaders? evidence from stakeholder commitments. *Journal of Financial Economics*, 127(3):519–545.

Pinker, S. (2011). Representations and decision rules in the theory of self-deception. *Behavioral and Brain Sciences*, 34(1):35–37.

Price, P. and Stone, E. (2004). Intuitive evaluation of likelihood judgment producers. *Journal of Behavioral Decision Making*, 17:39–57.

Probst, G. and Raisch, S. (2005). Organizational crisis: The logic of failure. *Academy of Management Perspectives*, 19(1):90–105.

Puri, M. and Robinson, D. (2007). Optimism and economic choice. *Journal of Financial Economics*, 86(1):71–99.

Quattrone, G. A. and Tversky, A. (1984). Causal versus diagnostic contingencies: On self-deception and on the voter's illusion. *Journal of personality and social psychology*, 46(2):237.

Rabin, M. and Schrag, J. L. (1999). First impressions matter: A model of confirmatory bias. *The Quarterly Journal of Economics*, 114(1):37–82.

Reuben, E., Rey-Biel, P., Sapienza, P., and Zingales, L. (2012). The emergence of male leadership in competitive environments. *Psychology, Public, Policy and Law*, 83(1):111–117.

Rivas, M. F. and Sutter, M. (2011). The benefits of voluntary leadership in experimental public goods games. *Economics Letters*, 112(2):176–178.

Rosenbaum, S. M., Billinger, S., and Stieglitz, N. (2014). Let's be honest: A review of experimental evidence of honesty and truth-telling. *Journal of Economic Psychology*, 45:181–196.

Saucet, C. and Villeval, M. C. (2018). Motivated memory in dictator games.

Saucier, D. A., Miller, C. T., and Doucet, N. (2005). Differences in helping whites and blacks: A meta-analysis. *Personality and Social Psychology Review*, 9(1):2–16.

Savin-Williams, R. (1979). Dominance hierarchies in groups of early adolescents. *Child development*, pages 923–935.

Schwardman, P. and van der Weele, J. (2019). Deception and self-deception. *Nature Human Behavior (in Press)*.

Schweitzer, M. E., Hershey, J. C., and Bradlow, E. T. (2006). Promises and lies: Restoring violated trust. *Organizational behavior and human decision processes*, 101(1):1–19.

Sedikides, C. and Green, J. D. (2009). Memory as a self-protective mechanism. *Social and Personality Psychology Compass*, 3(6):1055–1068.

Sedikides, C., Horton, R. S., and Gregg, A. P. (2007). The why's the limit: curtailing self-enhancement with explanatory introspection. *Journal of Personality*, 75(4):783–824.

Shalvi, S., Dana, J., Handgraaf, M. J., and De Dreu, C. K. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, 115(2):181–190.

Shamir, B., House, R. J., and Arthur, M. B. (1993). The Motivational Effects of Charismatic Leadership: A Self-Concept Based Theory. *Organization Science*, 4(4):577–594.

Sharot, T. and Garrett, N. (2016). Forming beliefs: Why valence matters. *Trends in cognitive sciences*, 20(1):25–33.

Sharot, T., Kanai, R., Marston, D., Korn, C. W., Rees, G., and Dolan, R. J. (2012). Selectively altering belief formation in the human brain. *Proceedings of the National Academy of Sciences*, 109(42):17058–17062.

Sharot, T., Korn, C. W., and Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature neuroscience*, 14(11):1475.

Shepperd, J. A. and Arkin, R. M. (1989). Determinants of self-handicapping: Task importance and the effects of preexisting handicaps on self-generated handicaps. *Personality and Social Psychology Bulletin*, 15(1):101–112.

Shipman, A. S. and Mumford, M. D. (2011). When confidence is detrimental: Influence of overconfidence on leadership effectiveness. *The Leadership Quarterly*, 22(4):649–665.

Shu, L. L. and Gino, F. (2012). Sweeping dishonesty under the rug: How unethical actions lead to forgetting of moral rules. *Journal of Personality and Social Psychology*, 102(6):1164.

Smith, M. K., Trivers, R., and von Hippel, W. (2017). Self-deception facilitates interpersonal persuasion. *Journal of Economic Psychology*, 63:93–101.

Snyder, M. L., Kleck, R. E., Strenta, A., and Mentzer, S. J. (1979). Avoidance of the handicapped: an attributional ambiguity analysis. *Journal of personality and social psychology*, 37(12):2297.

Soldà, A., Ke, C., Page, L., and von Hippel, W. (2019). Strategically delusional. *GATE WP No. 1908*.

Spence, S. A., Kaylor-Hughes, C., Farrow, T. F., and Wilkinson, I. D. (2008). Speaking of secrets and lies: the contribution of ventrolateral prefrontal cortex to vocal deception. *Neuroimage*, 40(3):1411–1418.

Steinel, W. and De Dreu, C. K. (2004). Social motives and strategic misrepresentation in social decision making. *Journal of personality and social psychology*, 86(3):419.

Story, A. L. (1998). Self-esteem and memory for favorable and unfavorable personality feedback. *Personality and Social Psychology Bulletin*, 24(1):51–64.

Stulp, G., Buunk, A., Verhulst, S., and Pollet, T. (2013). Tall claims? Sense and nonsense about the importance of height of US presidents. *The Leadership Quarterly*, 24(1):159–171.

Sullivan, P. S., Lansky, A., Drake, A., Investigators, H.-., et al. (2004). Failure to return for hiv test results among persons at high risk for hiv infection: results from a multistate interview project. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 35(5):511–518.

Sunstein, C. R., Bobadilla-Suarez, S., Lazzaro, S. C., and Sharot, T. (2016). How people update beliefs about climate change: Good news and bad news. *Cornell L. Rev.*, 102:1431.

Sutter, M. (2008). Deception through telling the truth?! experimental evidence from individuals and teams. *The Economic Journal*, 119(534):47–60.

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta psychologica*, 47(2):143–148.

Sweeny, K., Melnyk, D., Miller, W., and Shepperd, J. A. (2010). Information avoidance: Who, what, when, and why. *Review of general psychology*, 14(4):340–353.

Swift, S. A. and Moore, D. (2012). Bluffing, agonism, and the role of overconfidence in negotiation. In *The Oxford handbook of economic conflict resolution*, page 266. Oxford University Press.

Taylor, S. E. and Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological bulletin*, 103(2):193.

Taylor, S. E., Kemeny, M. E., Reed, G. M., Bower, J. E., and Gruenewald, T. L. (2000). Psychological resources, positive illusions, and health. *American psychologist*, 55(1):99.

Thompson, L. and Loewenstein, G. (1992). Egocentric interpretations of fairness and interpersonal conflict. *Organizational Behavior and Human Decision Processes*, 51(2):176–197.

Thornton, R. L. (2008). The demand for, and impact of, learning hiv status. *American Economic Review*, 98(5):1829–63.

Tigue, C. C., Borak, D. J., O'Connor, J. J., Schandl, C., and Feinberg, D. R. (2012). Voice pitch influences voting behavior. *Evolution and Human Behavior*, 33(3):210–216.

Todorov, A., Olivola, C. Y., Dotsch, R., and Mende-Siedlecki, P. (2015). Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. *Annual Review of Psychology*, 66(1):519–545.

Trivers, R. (1976). *Foreword in: The Selfish Gene, R. Dawkins*. Oxford University Press.

Trivers, R. (1985). Deceit and self-deception. in: Social evolution. *Benjamin/Cummings*, page 395–420.

Trivers, R. (2000). The elements of a scientific theory of self-deception. *Annals of the New York Academy of Sciences*, 907(1):114–131.

Trivers, R. (2010). Deceit and self-deception. In *Mind the gap*, pages 373–393. Springer.

Trope, Y., Gervey, B., and Bolger, N. (2003). The role of perceived control in overcoming defensive self-evaluation. *Journal of Experimental Social Psychology*, 39(5):407–419.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131.

van Dolder, D., van den Assem, M. J., Camerer, C. F., and Thaler, R. H. (2015). Standing United or Falling Divided? High Stakes Bargaining in a TV Game Show. *American Economic Review*, 105(5):402–407.

Van Swol, L. M. and Sniezek, J. A. (2005). Factors affecting the acceptance of expert advice. *British Journal of Social Psychology*, 44(3):443–461.

Vialle, I., Santos-Pinto, L. P., and Rulliere, J.-L. (2011). Self-Confidence and Teamwork: An Experimental Test. *SSRN Electronic Journal*.

von Hippel, W., Baker, E., Wilson, R., Brin, L., and Page, L. (2016). Detecting deceptive behaviour after the fact. *British Journal of Social Psychology*, 55(2):195–205.

Von Hippel, W., Lakin, J. L., and Shakarchi, R. J. (2005). Individual differences in motivated social cognition: The case of self-serving information processing. *Personality and Social Psychology Bulletin*, 31(10):1347–1357.

von Hippel, W. and Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences*, 34(01):1–16.

Von Neumann, J. and Morgenstern, O. (1953). *Theory of games and economic behavior*. Princeton University Press Princeton, NJ.

Vrij, A. (2004). Why professionals fail to catch liars and how they can improve. *Legal and criminological psychology*, 9(2):159–181.

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities*. John Wiley and Sons.

Vrij, A., Fisher, R., Mann, S., and Leal, S. (2006). Detecting deception by manipulating cognitive load. *Trends in cognitive sciences*, 10(4):141–142.

Vrij, A., Fisher, R. P., and Blank, H. (2017). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, 22(1):1–21.

Vrij, A. and Ganis, G. (2014). Theories in deception and lie detection. In *Credibility Assessment*, pages 301–374. Elsevier.

Vrij, A. and Mann, S. (2005). Police use of nonverbal behavior as indicators of deception. *Applications of nonverbal communication, ed. RE Riggio & RS Feldman*, pages 63–94.

Wang, J. T.-y., Spezio, M., and Camerer, C. F. (2010). Pinocchio's pupil: using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games. *American Economic Review*, 100(3):984–1007.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly journal of experimental psychology*, 20(3):273–281.

Watson, A. (2006). Self-deception and survival: Mental coping strategies on the western front, 1914-18. *Journal of Contemporary History*, 41(2):247–268.

Weinberg, B. A. (2009). A model of overconfidence. *Pacific Economic Review*, 14(4):502–515.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of personality and social psychology*, 39(5):806.

White, S. and Neale, M. A. (1994). The role of negotiator aspirations and settlement expectancies in bargaining outcomes. *Organizational Behavior and Human Decision Processes*, 57(2):303–317.

Wilson, T. D. and LaFleur, S. J. (1995). Knowing what you'll do: Effects of analyzing reasons on self-prediction. *Journal of Personality and Social Psychology*, 68(1):21.

Wilson, T. D., Wheatley, T. P., Kurtz, J. L., Dunn, E. W., and Gilbert, D. T. (2004). When to fire: Anticipatory versus postevent reconstrual of uncontrollable events. *Personality and Social Psychology Bulletin*, 30(3):340–351.

Wiswall, M. and Zafar, B. (2015). How do college students respond to public information about earnings? *Journal of Human Capital*, 9(2):117–169.

Yaniv, I., Benador, D., and Sagi, M. (2004). On not wanting to know and not wanting to inform others: Choices regarding predictive genetic testing. *Risk Decision and Policy*, 9(4):317–336.

Zimmerman, B. J., Moylan, A., Hudesman, J., White, N., and Flugman, B. (2011). Enhancing self-reflection and mathematics achievement of at-risk urban technical college students. *Psychological Test and Assessment Modeling*, 53(1):141–160.

Zimmermann, F. et al. (2019). The dynamics of motivated beliefs. Technical report, University of Bonn and University of Mannheim, Germany.

Zuckerman, M., DePaulo, B. M., and Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In *Advances in experimental social psychology*, volume 14, pages 1–59. Elsevier.

Appendix A

Appendix of Chapter 2

A.1 Summary Statistics

Table A.1 presents the mean values (with standard errors in parentheses) of all the main variables of interest across treatments and conditions, as well as the difference between *Persuasion-first* and *Accuracy-first* within an information condition. Table A.2 displays the mean values for the same variables for Study 2, as well as for the pooled data for *SCI* condition from both studies.

Table A.1: A summary of results in Study 1.

	NI			*GI*			SCI			Pooled (*GI*+*NI*+*SCI*)		
	Acc.1st	*Per.1st*	Δ	*Acc.1st*	*Per.1st*	Δ	*Acc.1st*	*Per.1st*	Δ	*Acc.1st*	*Per.1st*	Δ
Performance	24.28	23.33	**−0.95**	23.73	24.05	**0.32**	23.78	23.92	**0.14**	23.93	23.77	**-0.16**
	(0.423)	(0.511)	(0.662)	(0.492)	(0.468)	(0.679)	(0.411)	(0.425)	(0.591)	(0.255)	(0.270)	(0.372)
Belief about perf.	23.26	22.91	**−0.35**	21.92	22.91	**0.99***	22.35	23.67	**1.32***	22.51	23.16	**0.65****
	(0.556)	(0.602)	(0.819)	(0.639)	(0.736)	(0.975)	(0.557)	(0.485)	(0.741)	(0.338)	(0.355)	(0.490)
Overconfidence	-1.02	-0.41	**0.61**	-1.81	-1.15	**0.67**	-1.43	-0.25	**1.18*****	-1.42	-0.60	**0.82*****
	(0.403)	(0.537)	(0.669)	(0.435)	(0.503)	(0.665)	(0.373)	(0.358)	(0.517)	(0.233)	(0.272)	(0.358)
Belief about relative perf.	69.12%	71.46%	**2.34**%	69.96%	73.44%	**3.48**%	70.31%	73.74%	**3.43**%	69.80%	72.88%	**3.08**%**
	(1.927)	(1.917)	(2.719)	(2.005)	(2.004)	(2.835)	(1.841)	(1.686)	(2.500)	(1.107)	(1.079)	(1.546)
Overplacement	16.33%	23.85%	**7.52****%	19.64%	21.81%	**2.17**%	21.57%	24.37%	**2.80**%	19.19%	23.34%	**4.15**%**
	(2.444)	(2.611)	(3.574)	(2.450)	(2.481)	(3.487)	(2.367)	(2.406)	(3.375)	(1.398)	(1.440)	(2.006)
Feedback	–	–	–	77.92%	80%	**2.08**%	67.4%	79.69%	**12.29**%***	–	–	–
				(1.801)	(1.891)	(2.611)	(2.106)	(2.075)	(2.958)			
Feedback bias	–	–	–	-1.18%	-0.17%	**1.01**%	-11.87%	-0.03%	**11.84**%***	–	–	–
				(0.921)	(1.011)	(1.367)	(1.885)	(1.923)	(2.692)			
Av. guessed score	21.81	21.94	**0.13**	21.85	22.38	**0.53**	21.79	22.77	**0.98***	21.82	22.36	**0.55****
	(0.459)	(0.504)	(0.681)	(0.446)	(0.475)	(0.652)	(0.426)	(0.408)	(0.590)	(0.255)	(0.268)	(0.370)
Reviewers' bias	-2.47	-1.39	**1.07**	-1.88	-1.67	**0.21**	-1.99	-1.14	**0.85**	-2.11	-1.40	**0.71**
	(0.549)	(0.579)	(0.798)	(0.543)	(0.510)	(0.745)	(0.490)	(0.533)	(0.723)	(0.304)	(0.312)	(0.435)
Av. convincingness	3.09	2.93	**−0.16**	3.21	3.26	**0.05**	3.10	3.29	**0.19***	3.13	3.16	**0.03**
	(0.081)	(0.103)	(0.131)	(0.073)	(0.077)	(0.106)	(0.081)	(0.074)	(0.110)	(0.045)	(0.050)	(0.068)
Obs.	98	97	195	96	96	192	100	97	197	294	290	583

Note: Table A.1 reports mean values of all measures (with standard errors in parentheses), and two-sided Mann-Whitney tests between *Accuracy-first* and *Persuasion-first* treatments within each information condition, and pooled across all information conditions (*NI*, *GI*, *SCI*). *** p <0.01, ** p <0.05, * p <0.10.

Table A.2: A summary of results in Study 2 with pooled data on *SCI* condition.

	SCI (lab)			*SCI* (MTurk + lab)		
	Acc.1st	Per.1st	Δ	Acc.1st	Per.1st	Δ
Performance	20.02	18.92	**−1.1**	22.53	22.22	**−0.31**
	(0.682)	(0.585)	(0.899)	(0.383)	(0.395)	(0.550)
Belief about perf.	18.46	20.18	**1.72**	21.05	22.48	**1.43****
	(0.833)	(0.748)	(1.119)	(0.486)	(0.430)	(0.650)
Overconfidence	-1.56	1.26	**2.82*****	-1.47	0.27	**1.74*****
	(0.491)	(0.519)	(0.714)	(0.297)	(0.300)	(0.422)
Belief about relative perf.	59.22%	65.98%	**6.76**%	66.61%	71.10%	**4.49**%*
	(2.937)	(2.341)	(3.756)	(1.622)	(1.397)	(2.144)
Overplacement	4.22%	10.98%	**6.76**%	15.79%	19.81%	**4.03**%
	(4.889)	(4.686)	(6.772)	(2.356)	(2.301)	(3.294)
Feedback	62.2%	75%	**12.8**%**	65.67%	78.10%	**12.43**%***
	(3.783)	(2.931)	(4.786)	(1.891)	(1.698)	(2.544)
Feedback bias	-4.53%	11.93%	**16.47**%***	-9.42%	4.04%	**13.46**%***
	(3.649)	(3.386)	(4.978)	(1.764)	(1.770)	(2.499)
Av. guessed score	22.02	23.07	**1.05**	21.87	22.87	**1.00****
	(0.578)	(0.635)	(0.858)	(0.342)	(0.344)	(0.485)
Reviewers' bias	2.00	4.15	**2.15**	-0.661	0.656	**1.32****
	(0.634)	(0.770)	(0.997)	(0.417)	(0.484)	(0.638)
Av. convincingness	3.62	3.43	**−0.18**	3.27	3.34	**0.07**
	(0.088)	(0.123)	(0.151)	(0.064)	(0.064)	(0.091)
N	50	50	100	150	147	297

Note: Table A.2 reports mean values of all measures (with standard errors in parentheses), and two-sided Mann-Whitney tests between *Accuracy-first* and *Persuasion-first* treatments for all measures related to performance, beliefs, and feedback. We also report the estimated average guessed score, reviewers' bias in their guess, and their ratings of convincingness using OLS regressions with robust standard errors clustered by session. (For all measures from the reviewers', MW tests do not apply anymore since in Study 2 reviewers rated multiple essays and one essay was reviewed by 5 different reviewers. Hence, we do not have fully independent observations.) *** p <0.01, ** p <0.05, * p <0.10.

A.2 Additional Analyses

A.2.1 Individual Characteristics

Table A.3: Individual characteristics, by treatment

	Number of individuals	Mean score	Females (Percentage)	Mean age	Mean OCQ
	(1)	(2)	(3)	(4)	(5)
NI x Acc. 1st	98	24.28	0.480	35.92	4.12
NI x Pers. 1st	97	23.33	0.406	34.05	5.79
GI x Acc. 1st	96	23.73	0.479	36.29	4.54
GI x Pers. 1st	96	24.05	0.442	35.41	4.05
SCI x Acc. 1st	100	23.78	0.745	35.96	4.30
SCI x Pers. 1st	97	23.92	0.515	35.86	4.20
SCI lab x Acc. 1st	50	20.02	0.495	22.7	6.12
SCI lab x Pers. 1st	50	18.92	0.495	21.2	6.36

Note: A one-way between-subject ANOVA shows that there are no significant difference in the distribution of participants in terms of performance between treatments in Study 1 ($F(5, 578) = 0.50$; $p = 0.779$). We also find no significant difference in participants' performance between treatments in Study 2 (two tailed t-test: $p = 0.224$; Kolmogorov-Smirnov test: $p = 0.711$). A Chi2 test showed that there is no significant difference in the proportion of male and female between treatments in Study 1 ($Chi2(5) = 2.360$; $p = 0.797$). We also find no significant difference in the distribution of gender Study 2 (two-sample test of proportion: $p = 0.317$). Another one-way between-subject ANOVA shows that there are no significant difference in the distribution of participants in terms of age between treatments in Study 1 ($F(5, 578) = 0.55$; $p = 0.740$) and between treatments in Study 2 (two tailed t-test: $p = 0.100$; Kolmogorov-Smirnov test: $p = 0.544$). Finally, a one-way between-subject ANOVA shows that there are no significant difference in the distribution of participants in terms of dispositional overconfidence measured by the OCQ between treatments in Study 1 ($F(5, 578) = 1.59$; $p = 0.161$) and between treatments in Study 2 (two tailed t-test: $p = 0.816$; Kolmogorov-Smirnov test: $p = 0.393$).

A.2.2 Effect of Anticipation of Strategic Interactions on Beliefs

Table A.4 presents the determinants of participants beliefs about their performance. Columns (1), (3), (6), (9) and (12) report the OLS regression of participants' beliefs about their performance on the treatment dummy "Persuasion", controlling for performance. Columns (4), (7), (10) and (13) report the OLS regression of participants' beliefs about their performance on the treatment dummy and the proportion of correct answers contained in the feedback. In Columns (2), (5), (8), (11) and (14), we further control for participants individual characteristics (sex, age and OCQ). Table A.5 presents the same models with participants' beliefs about their relative performance as the dependent variable.

Table A.4: Determinants of Participants' Beliefs about Performance.

	Study 1								Study 2			Pooled		
Dep. Var: Beliefs	*NI*		*GI*			*SCI*			*SCI* (lab)			*SCI* (MTurk +lab)		
about perf.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Feedback	–	–	–	0.285***	0.282***	–	0.094***	0.087***	–	0.041*	0.045*	–	0.080***	0.071***
				(0.018)	(0.018)	(0.017)	(0.016)		(0.023)	(0.024)		(0.014)	(0.014)	
Persuasion	0.376	0.201	0.650	0.395	0.401	1.196**	0.167	0.185	2.776***	1.192	0.821	1.709***	0.429	0.551
	(0.654)	(0.652)	(0.666)	(0.631)	(0.638)	(0.516)	(0.719)	(0.692)	(0.722)	(1.147)	(1.158)	(0.419)	(0.642)	(0.619)
Performance	0.752***	0.696***	1.052***	–	–	0.903***	–	–	0.960***	–	–	0.904***		
	(0.071)	(0.073)	(0.071)			(0.062)			(0.081)			(0.044)		
Female	–	-2.304***	–	–	-0.430	–	–	-1.639**	–	–	0.278	–	–	-1.089*
		(0.651)			(0.673)			(0.673)			(1.118)			(0.603)
Age	–	0.029	–	–	0.038	–	–	0.042	–	–	-0.254**	–	–	0.087***
		(0.030)			(0.030)			(0.033)			(0.125)			(0.028)
OCQ	–	-0.015	–	–	-0.002 –	–		-0.219***	–	–	-0.101	–	–	-0.203***
		(0.064)			(0.069)			(0.063)			(0.108)			(0.057)
Constant	4.988***	6.487***	-3.041*	-0.329	-1.228	0.886	16.027***	16.747***	-0.757	15.894***	21.915***	0.696	15.768***	15.270***
	(1.773)	(2.209)	(1.753)	(1.435)	(1.840)	(1.529)	(1.223)	(1.699)	(1.691)	(1.652)	(3.413)	(1.042	(1.025)	(1.377)
R-squared	0.409	0.526	0.539	0.587	0.591	0.526	0.154	0.238	0.603	0.054	0.100	0.592	0.114	0.193
Observations	194	193	192	192	191	197	197	196	100	100	100	297	297	296

Notes: Table A.4 reports OLS regressions with standard errors in parentheses. Columns (1) to (8) shows the results for the observations of participants in Study 1. Columns (9) and (11) shows the results for the observations of participants in Study 2. Columns (12) and (14) shows the results for the observations of participants in studies 1 and 2 pooled together (*SCI* treatment only). *** p <0.01, ** p <0.05, * p <0.10.

Table A.5: Determinants of Participants' Beliefs about Relative Performance.

	Study 1								Study 2			Pooled		
Dep. Var: Beliefs	*NI*		*GI*			*SCI*			*SCI* (lab)			*SCI* (MTurk +lab)		
about relative perf.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Feedback	–	–	–	0.665***	0.631***	–	0.233***	0.204***	–	0.152*	0.162**	–	0.221***	0.190***
				(0.062)	(0.065)		(0.058)	(0.058)		(0.078)	(0.081)		(0.047)	(0.048)
Persuasion	4.593**	3.779*	2.569	2.0.94	2.271	3.099	0.563	0.962	9.175**	4.814	4.474	5.194***	1.747	2.120
	(2.245)	(2.247)	(2.149)	(2.250)	(2.230)	(2.055)	(2.513)	(2.448)	(3.237)	(3.837)	(3.959)	(1.741)	(2.155)	(2.132)
Performance	2.344***	2.185***	2.735***	–	–	2.423***	–	–	2.195***	–	–	2.284***		
	(0.242)	(0.253)	(0.230)			(0.249)			(0.361)			(0.184)		
Female	–	-7.449***	–	–	-4.556**	–	–	-7.470***	–	–	1.541	–	–	-4.871**
		(2.289)			(2.277)			(2.383)			(3.822)			(2.076)
Age	–	-0.012	–	–	0.232**	–	–	-0.039	–	–	-0.237	–	–	0.155
		(0.103)			(0.104)			(0.118)			(0.427)			(0.096)
OCQ	–	-0.009	–	–	-0.180	–	–	-0.580**	–	–	0.070	–	–	-0.454**
		(0.219)			(0.241)			(0.223)			(0.371)			(0.195)
Constant	12.223**	20.109***	5.061	18.144***	15.393**	12.701**	54.576***	64.125**	15.271**	49.763***	53.276***	15.161***	52.129***	53.954***
	(6.086)	(7.609)	(5.654)	(5.117)	(6.433)	(6.095)	(4.279)	(6.014)	(7.579)	(5.524)	(11.670)	(4.326)	(3.439)	(4.738)
R-squared	0.332	0.365	0.433	0.380	0.409	0.334	0.085	0.163	0.299	0.068	0.073	0.353	0.082	0.124
Observations	194	193	192	192	191	197	197	196	100	100	100	297	297	296

Notes: Table A.5 reports OLS regressions with standard errors in parentheses. Columns (1) to (8) shows the results for the observations of participants in Study 1. Columns (9) and (11) shows the results for the observations of participants in Study 2. Columns (12) and (14) shows the results for the observations of participants in studies 1 and 2 pooled together (*SCI* treatment only). *** p <0.01, ** p <0.05, * p <0.10.

Columns (1) and (2) present the results for the *NI* condition. We do not observe any significant effect of being in the *Persuasion-first* treatment in this condition. Higher performance always leads to higher beliefs about their absolute performance. Model (2) also shows that women tend to be less confident than men about their performance in the absence of feedback. Columns (3) to (5) from table A.4 present the results of the estimations for the *GI* condition. These models show that being in the *Persuasion-first* treatment affects positively participants' beliefs about their score but the effect is not significant, after controlling for performance. These models also show that the feedback has a significant effect on beliefs at the 1% level. Finally, columns (6) to (14) present the results for the *SCI* condition across two studies. Controlling for performance in the baseline model, we observe that being in the *Persuasion-first* treatment leads to higher beliefs about their absolute performance and the effect is significant at the 1% level. However, when we add the Feedback as an explanatory variable, the treatment effect disappears and the feedback has a positive effect on beliefs and the effect is significant at the 1% level in Study 1 and when pooling Study 1 and 2, but is only significant at the 10% level in Study 2. The difference in the level of significance may be driven by very different number of observations we have in each study. Table A.5 displays similar (stronger) results when we use participants' beliefs about their rank as the dependent variable.

A.2.3 Determinants of Information Sampling and its Impact on Beliefs about Performance and Relative Performance

Table A.6: Determinants of Information sampling.

	Study 1				Study 2		pooled	
Dep. Var:	*GI*		*SCI*		*SCI* (lab)		*SCI* (MTurk + lab)	
Feedback	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Per. 1st	1.025	0.860	11.973***	12.319***	14.250***	14.499***	12.978***	13.073***
	(1.371)	(1.381)	(2.631)	(2.631)	(4.696)	(4.791)	(2.354)	(2.350)
Performance	3.277***	3.231***	2.312***	2.427***	1.318**	1.333**	1.778***	1.782***
	(0.146)	(0.154)	(0.319)	(0.350)	(0.524)	(0.530)	(0.249)	(0.278)
Female	–	0.226	–	-2.105	–	-6.860	–	-4.344*
		(1.412)		(2.719)		(4.682)		(2.389)
Age	–	-0.045	–	-0.086	–	0.587	–	-0.050
		(0.065)		(0.136)		(0.533)		(0.117)
OCQ	–	-0.252*	–	0.355	–	-0.160	–	0.109
		(0.148)		(0.266)		(0.460)		(0.232)
Constant	0.154	3.903	12.425	11.995	35.814***	26.880	25.609***	28.525***
	(3.608)	(4.294)	(7.802)	(9.441)	(10.993)	(18.342)	(5.848)	(6.883)
R-squared	0.727	0.732	0.277	0.291	0.125	0.157	0.212	0.224
Observations	192	191	197	196	100	100	297	296

Notes: Table A.6 reports OLS regressions with standard errors in parentheses. We consider the observations of participants in the *GI* and *SCI* treatment only. Columns (1) and (2) shows the results for the observations of participants in *GI*. Columns (3) and (4) shows the results for the observations of participants in *SCI* from Study 1. Columns (5) and (6) shows the results for the observations of participants in *SCI* from Study 2. Columns (7) and (8) shows the results for the observations of participants in *SCI* from both Study 1 and 2 pooled together. *** p <0.01, ** p <0.05, * p <0.10.

Table A.6 presents the determinants of the proportion of correct answers revealed through the feedback in *Given Information* and *Self-Chosen Information* conditions. Columns (1), (3), (5), and (7) report the OLS regressions of the proportion of correct answers revealed through the feedback on treatment dummy and performance. We control for participants' characteristics (sex, age and OCQ) in columns (2), (4), (6) and (8). Columns (1) and (2) shows the results for the observations of participants in *Given Information*. Columns (3) and (4) shows the results for the observations of participants in *Self-Chosen Information* from Study 1. Columns (5) and (6) shows the results for the observations of participants in *Self-Chosen Information* from Study 2. Columns (7) and (8) shows the results for the observations of participants in *Self-Chosen Information* from both Study 1 and 2 pooled together.

As expected, models (1) and (2) shows no treatment effect on the feedback content since the feedback is exogenous in the *Given Information* condition. On the

other hand, models (3) to (8) show a positive and strongly significant (1%) treatment effect on the proportion of correct answers revealed through the feedback.

Table A.7: Effect of information sampling on beliefs about performance and relative performance.

Dep. Var:	*SCI* (MTurk)		*SCI* (lab)		*SCI* (MTurk +lab)	
Beliefs about	perf.	relative perf.	perf.	relative perf.	perf.	relative perf.
	(1)	(2)	(3)	(4)	(5)	(6)
Feedback	0.091**	0.251	0.184**	0.660**	0.125	0.384***
	(0.044)	(0.170)	(0.077)	(0.295)	(0.038)	(0.146)
Performance	0.677***	1.807***	0.701***	1.363**	0.702***	1.694***
	(0.130)	(0.503)	(0.570)	(0.149)	(0.088)	(0.340)
Female	-0.475	-4.134*	1.770	6.181	0.259	-1.189
	(0.561)	(2.175)	(1.265)	(4.833)	(0.532)	(2.047)
Age	-0.018	-0.203*	-0.209*	-0.239	-0.027	-0.133
	(0.028)	(0.109)	(0.124)	(0.475)	(0.025)	(0.095)
OCQ	-0.048	-0.116	-0.016	0.293	-0.035	-0.029
	(0.057)	(0.220)	(0.108)	(0.413)	(0.049)	(0.189)
Constant	1.271	20.369**	-3.152	-8.817	-1.985	8.328
	(2.061)	(7.997)	(5.366)	(20.504)	(1.978)	(7.615)
First-stage F-stat	15.61	15.61	3.51	3.51	16.70	16.70
Observations	196	196	100	100	296	296

Notes: Table A.7 reports 2SLS regressions with standard errors in parentheses. Feedback is instrumented by the treatment dummy. Columns (1) to (2) shows the results from Study 1. Columns (3) and (4) shows the results from Study 2. Columns (5) and (6) shows the results for both studies 1 and 2 pooled together. *** p <0.01, ** p <0.05, * p <0.10.

Table A.7 reports the corresponding regressions reported in Table 2.2 with all the control variabled included. Results stay robust.

A.2.4 Causal identification: the effect of confidence on persuasion (Robustness check)

Table A.8: Causal identification of the effect of participants' beliefs about performance and relative performance on persuasiveness.

Dep. Var:	*SCI* (MTurk)		*SCI* (lab)		*SCI* (MTurk + lab)	
Persuasiveness	(1)	(2)	(3)	(4)	(5)	(6)
Beliefs about perf.	0.936	–	0.407*	–	0.641***	–
	(0.664)		(0.239)		(0.201)	
Beliefs about	–	0.323	–	0.112**	–	0.209***
relative perf.		(0.271)		(0.057)		(0.071)
Feedback dummy	-0.362	-0.078	-0.908	-1.149	-0.270	-0.129
	(0.848)	(0.892)	(0.834)	(0.881)	(0.336)	(0.439)
Performance	-0.611	-0.553	-0.042	0.099	-0.337*	-0.248
	(0.599)	(0.662)	(0.208)	(0.130)	(0.201)	(0.212)
Female	0.644	1.532	-0.701	-0.668	0.053	0.5483
	(1.472)	(1.351)	(0.805)	(0.760)	(0.438)	(0.563)
Age	0.066	0.058	-0.118	-0.182	-0.032	-0.020
	(0.075)	(0.068)	(0.120)	(0.137)	(0.025)	(0.040)
OCQ	-0.057	-0.057	0.020	-0.017	-0.015	-0.032
	(0.080)	(0.077)	(0.045)	(0.054)	(0.029)	(0.031)
Constant	14.631***	9.350	18.661***	18.485***	17.097***	14.075***
	(3.056)	(7.285)	(2.108)	(2.817)	(1.104)	(1.725)
First-stage F-stat	37.64	18.85	144.10	9.98	136.79	65.21
Observations	196	196	100	100	296	296

Notes: Table A.8 reports 2SLS regressions with standard errors in parentheses. Participants' beliefs are instrumented by the treatment dummy. Columns (1) and (2) shows the results from Study 1. Columns (3) and (4) shows the results from Study 2. Columns (5) to (6) shows the results from studies 1 and 2 pooled together. *** $p<0.01$, ** $p<0.05$, * $p<0.10$.

A.2.5 Causal identification: the effect of confidence on persuasion (*Given Information* condition)

In the *Given Information* condition, the feedback given to the participants has a random component. The questions selected, for all, to be used as feedback, may have been easier or harder for different participants, relative to the other questions they faced. If the question chosen for feedback happen to have been relatively hard in comparison to the other question he/she faced, this contestant's feedback may look more negative than his/her result over the whole test. We use this exogenous variation in positive *vs.* negative feedback bias to study the effect of the induced variations in confidence on persuasiveness.

Table A.9: Effect of participants' beliefs about performance and relative performance on persuasiveness.

Dep Var:	*Given Information* (*GI*)			
Persuasiveness	(1)	(2)	(3)	(4)
Beliefs about perf.	0.312*	0.310*	–	–
	(0.171)	(0.165)		
Beliefs about	–	–	0.285	0.281
relative perf.			(0.197)	(0.187)
Feedback dummy	-0.494	-0.581	-1.852	-1.851
	(0.675)	(0.657)	(1.431)	(1.325)
Performance	0.039	0.070	-0.393	-0.331
	(0.190)	(0.181)	(0.530)	(0.484)
Female	–	1.121*	–	2.243**
		(0.613)		(1.119)
Age	–	-0.049*	–	-0.089*
		(0.065)		(0.084)
OCQ	–	0.016	–	0.043
		(0.065)		(0.084)
Constant	14.371***	14.936**	11.774***	12.537***
	(1.672)	(1.886)	(2.347)	(2.516)
First-stage F-statistic	99.43	49.20	53.39	28.18
Observations	192	191	192	191

Notes: Table A.9 reports 2SLS regressions for *GI* condition only with standard errors in parentheses. Participants' beliefs are instrumented by the Feedback variable. *** p <0.01, ** p <0.05, * p <0.10.

Table A.9 reports the corresponding 2SLS estimations of the effect of confidence on participants' persuasiveness, using the exogenous variation in feedback as an instrument. The dependent variable is participants' persuasiveness measured as the average reviewers' guess of the participants' score in the test. The independent variables include participants' beliefs about their absolute (or relative) performance instrumented by the exogenous variation in feedback, controlling for actual performance and whether feedback was mentioned in the essay in models (1) and (3), and further controlling for participants' individual characteristics (sex, age and OCQ) in models (2) and (4).

Table A.9 shows that an increase in beliefs has a positive effect on participants' level of persuasiveness. However, the effect is only (marginally) significant in models (1) and (2). Going back to Hypothesis 1, it is likely that the lack of effect

is due to the fact that participants do not have much room to inflate their beliefs in the *Given Information* condition. Indeed, one can think that it is harder to ignore an incorrect answer when it is displayed on the screen. If this hypothesis is true, we should find a stronger effect of beliefs on persuasion in the *Self-Chosen Information* condition. We have shown in Subsection 2.3.5 in the main text that this is indeed the case.

A.3 Essays content

A.3.1 Essays samples

Here is an example of an essay from a participant who was assigned to the *Self-Chosen Information* x *Persuasion-first* condition: *That test was a snap! I had the answers just rolling out of my head on to the answer list, usually before I even read the answer choices. I did the check to see if I got them right, I even tested most of the ones I was less sure about and got 9 of the 10 right! The other was a silly misclick, but I definetly knew the rest! I Even realized i hit the wrong thing as I hit next. I doubt many others did half as well as I did.* (actual score: 21/31).

In contrast, here is an example of an essay from a participant who was assigned to the *No Information* x *Persuasion-first* condition: *My friends have always joked with me that I would be enormously successful if I could find a way to profit from all the random information and bits of trivia that I know. I tend to have an excellent ability to remember seemingly insignificant details and almost never pass up an opportunity to learn something. Regardless of whether I'm reading a book, listening to a presentation, reading wikipedia or watching television, I'm always absorbing information. Perhaps this is my opportunity to finally earn some sort of return on my investment, I hope that you'll trust me when I say that I knew every question in the quiz that I just took.* (actual score: 28/31)

A.3.2 Summary Statistics

In this section, we investigate whether participants' essay content differ across treatment and information condition. We asked MTurk workers who did not participate in Study 1 (neither as a main participants nor as a reviewer) to code the content of each essay. Each essay was coded by 5 MTurk workers and each MTurk worker coded 18 essays. MTurk workers were informed that the authors of those essays undertook a knowledge test. They were also informed that some of the participants received a feedback about 10 questions of the test and were

told how many questions out of these 10 questions they answered correctly.

For each essay, the MTurk workers were asked to answer the following yes/no questions: (i) The participant mentioned the number (or percentage) of questions he or she thinks that he or she correctly answered during the test; (ii) The participant mentioned his or her rank compared to the other MTurkers (e.g., I think I did better than 80% of the other participants); (iii) The participant mentioned his or her own qualities or characteristics (e.g., I am good at history) and (iv) The participant mentioned the feedback he or she received about his or her performance. Table A.10 reports the average proportion of participants that mentioned the features summarised above in their essays.

Table A.10: Type of information mentioned in written Essays

Treatment-condition	number/percentage of correct answers	relative performance in the sample	own characteristics	feedback
NI x Acc. 1st	51.02%	17.35%	77.55%	–
NI x Per. 1st	50.52%	18.56%	77.32%	–
GI x Acc. 1st	65.63%	25.00%	64.58%	25.00%
GI x Per. 1st	65.00%	20.83%	80.21%**	46.88%**
SCI x Acc. 1st (MTurk)	56.00%	16.00%	77.00%	26.00%
SCI x Per. 1st (MTurk)	68.04%*	15.46%	70.10%	46.39%***
SCI x Acc. 1st (lab)	58.00%	18.00%	74.00%	42.00%
SCI x Per. 1st (lab)	68.00%	12.00%	78.00%	36.00%

Note: Table A.10 summarises the proportion of participants who mentioned a particular feature in their essay and the tests of proportions between *Accuracy-first* and *Persuasion-first* treatments within each information condition. *** p <0.01, ** p <0.05, * p <0.10.

Table A.10 shows that most participants (between 65-80%) wrote about their own characteristics (such as being good at Trivia or what their major is) and the scores or correctness rate they thought had in the test (between 50-68%). Only about 20% mentioned their beliefs about their relative position in the distribution. Overall, around 35% of the participants talked about the feedback they received in the *Given Information* and *Self-Chosen Information* conditions.

The identification assumption we made in Section 2.3.5 is that the treatment affects persuasiveness only through participants' beliefs. However, participants in the *Persuasion-first* treatment under *Self-Chosen Information* conditions could perhaps simply reveal more often the feedback they received in their essays (since they had more positive feedback than in the *Accuracy-first* treatment). Such a difference could potentially has an impact on persuasiveness independent from participants' confidence. We indeed observe that in both the *Given Information*

condition and the *Self-Chosen Information* condition in Study 1, participants were more likely to mention their feedback in the *Persuasion-first* treatment than in the *Accuracy-first* treatment (46.88% vs. 25% in *Given Information* condition, and 46.39% vs. 26% in *Self-Chosen Information* condition.

A.3.3 Hedging Behavior

In the Persuasion Task, participants are incentivized (i) to get the higher score as possible from reviewers and (ii) to be as convincing as possible. Because participants were presented with both questions at the same time, there is a possibility that they will engage in hedging behavior. To examine this issue, we coded each essay according to (i) whether the participant mentioned his performance at all in the essay and (ii) the overall valence of the essay (positive vs. neutral vs. negative). Out of the 95% of participants who mentioned their performance in their essay, 92% of them wrote positively about their performance, 4% reported an "OK/average" (neutral) performance and only 4% talked negatively about their performance. Out of those who mentioned their performance, 18.86% mentioned a number explicitly. We compared the numbers claimed to the estimate of their absolute performance from the Accuracy Task, and find that 87.60% claimed a number greater or equal to this estimate. These numbers suggest that most participants did not try to claim a lower performance in order to be rated as more convincing by the reviewers.

A.4 Instructions

A.4.1 Instructions (MTurk) - Accuracy-first x Given Information

— Main participants —

PART 1

Please rate your familiarity with each item by selecting the appropriate number from 0 to 4. If you are very familiar with an item, choose the number 4. If you have never heard of an item, choose the number 0.

	0 Never heard of it	1	2 Somewhat familiar	3	4 Very familiar
Bay of Pigs	○	○	○	○	○
United States	○	○	○	○	○
euphemism	○	○	○	○	○
sentence stigma	○	○	○	○	○
Dale Carnegie	○	○	○	○	○
Gail Brennan	○	○	○	○	◉
Lewis Carroll	○	○	○	○	○
art deco	○	○	○	○	○
meta-toxins	○	○	○	○	○
consumer apparatus	○	○	○	○	○
hyperbole	○	○	○	○	○
Charlotte Bronte	○	○	○	○	○
Artemis	○	○	○	○	○
free will	○	○	○	○	○
shunt-word	○	○	○	○	○
myth	○	○	○	○	○
cholarine	○	○	○	○	○
a cappella	○	○	○	○	○
The Aeneid	○	○	○	○	○
double entendre	○	○	○	○	○
alliteration	○	○	○	○	○
Queen Shattuck	○	○	○	○	○
Houdini	○	○	○	○	○
Murphy's Last Ride	○	○	○	○	○
blank verse	○	○	○	○	○
Antigone	○	○	○	○	○

PART 2

The second part of this study involves a general knowledge test in the form of a Multiple Choice Questionnaire.

The test is composed of 31 questions. You will have 15 seconds per questions to make your decision. Once you have made your decision, press the '>>' button to start the next question. If you fail to select an answer before the end of the 15 seconds, the next question will start automatically.

WARNING: You have to press the '>>' button to enter your answer. if you don't click on the '>>' button, the computer will score it as if you did not answer this question.

Click the '>>' button to start the practice trial before the real test.

PART 3

Now that you completed the knowledge test, you will have the chance to earn up to 2 extra dollars.

For this next task we would now like you to estimate how well you did on the general knowledge test. The closer your estimate is to your true performance, the more money you will earn.

In order to help you make this judgement, we will show you 10 of the questions and indicate whether you answered them correctly or not.

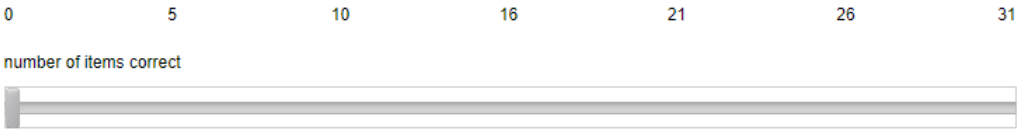
Press the '>>' button to see the answers.

Elicitation 1

You now have the opportunity to earn additional money by estimating your performance on the test. You will be paid as a function of the accuracy of your estimate.

You will earn $0.50 on this task if your estimate is in the 50% of most accurate estimates. You will earn another $0.50 if your estimate is in the 10% of most accurate estimates.

Please estimate how many questions that you think you scored correctly on the general knowledge



>>

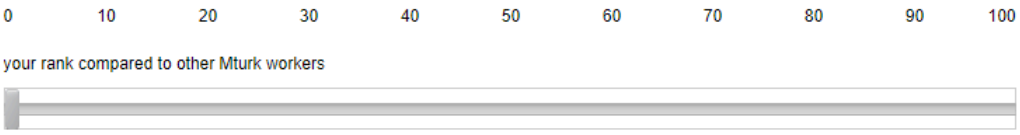
Elicitation 2

Thinking about your performance again, how do you think you compare to the average Mturk worker on this task, *in terms of number of items correct?*

On the slider below, place yourself between 0% and 100%.

For example, if you think you did better than 3/4 of your fellow MTurk users but worse than 1/4 of them, you will choose 75%. Conversely, if you think that you did better than 1/4 of your fellow MTurk users but worse than 3/4, you will choose 25%.

You will earn $0.50 on this task if the estimate of your rank compared to your fellow MTurk users is in the 50% of most accurate estimates. You will earn an additional $0.50 if the estimate of your rank is in the 10% of most accurate estimates.



>>

PART 4

You now have the opportunity to earn up to an additional $2 by writing a short essay to convince some of your fellow MTurk workers about how well you performed on the test.

Your payment for this task will depend **both** on (A) how well the reviewers think you performed on the test and (B) on how convincing your essay is.

(A) If your score is rated in the 50% best scores by the reviewers, you will be paid $0.50 and an additional $0.50 if your score is rated in the 10% best scores.

(B) If your essay is rated in the 50% of the most convincing essays by the reviewers, you will get $0.50 and if your essay is rated in the 10% of most convincing essays, you will earn another $0.50, independent of how high they think you scored.

We will not be providing your fellow MTurk workers with any information about how you actually did, so you need to put your best foot forward in an effort to be as convincing as possible about your performance.

When your essay is done, please hit the '>>' button.

>>

Final screen

Thank you for your participation.

You will receive your earnings for your participation and for the accuracy of your estimate within 3 days. You will receive your earnings for how convincing you were as soon as we collect the reviewers' evaluation (which may take a few weeks).

— Reviewers —

In a previous study, your fellow MTurk workers took a 31 questions knowledge test. At the end of the test, we ask them to write an essay about their performance.

Your task during this experiment will be to rate one of these short essays both on (A) how convincing you think the essay is and (B) how many questions you think the author answered correctly in the knowledge test.

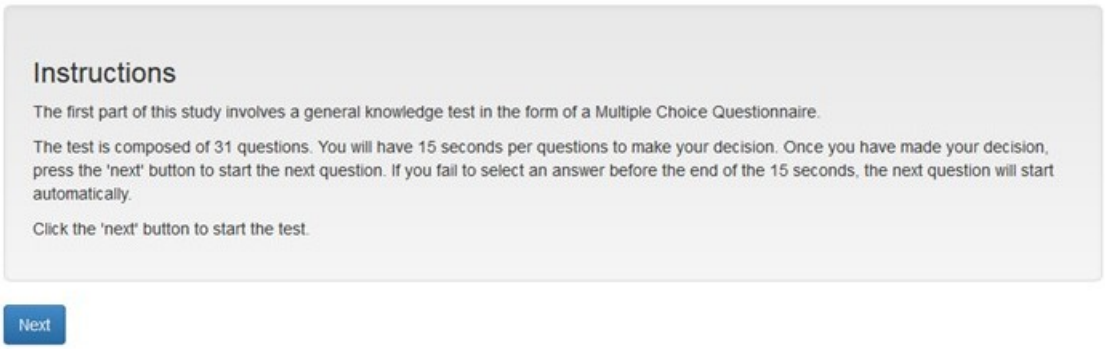
When you are ready to start, please hit the '>>' button.

A.4.2 Instructions (lab) - Persuasion-first x *Self-Chosen Information*

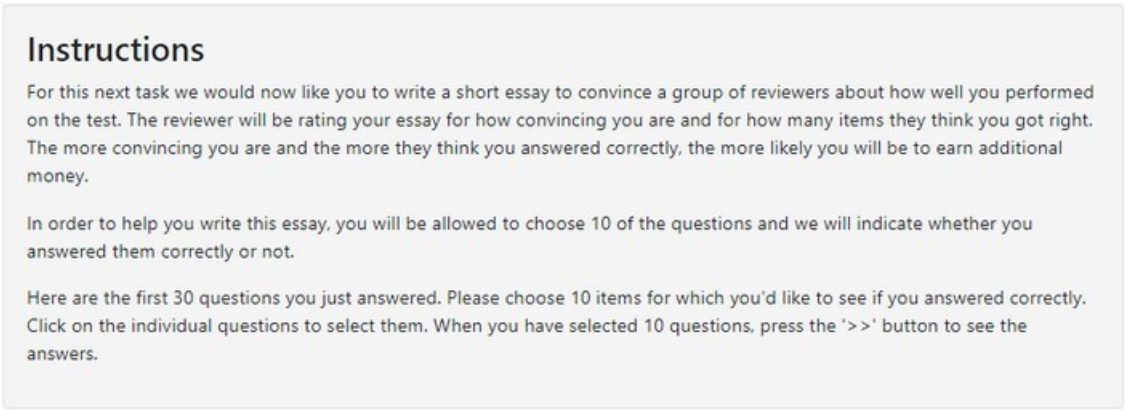
Note: All the instructions are displayed on the participants' screens.

— Main participants —

PART 1



PART 2



Select the items you want to check:

- ☐ What bird has the widest wingspan?
- ☐ The unit of electrical resistance was named after whom?
- ☐ Titan is a moon of which planet?
- ☐ How many pieces are on a chessboard at the start of a game?
- ☐ Which of these is the largest in area?
- ☐ In Greek mythology, who was the multi-headed dog, encircled by a serpent, that guarded the portal to the underworld?
- ☐ At the opening ceremony of every Olympic Games when the athletes parade into the stadium, what is traditionally the first nation to enter?

☐ Which of these types of music did not originate in the Caribbean?

☐ What is the name of the engraved stone, discovered in 1799, that provided a key to deciphering the languages of ancient Egypt?

☐ The Dalai Lama is a high lama in which religion?

☐ Which Scotsman took out a patent in 1876 that was the nucleus of the telephone?

☐ What is another name for a blood clot?

☐ Which of these countries is not landlocked?

☐ The tibia and fibula are found where in the human body?

☐ ”Facebook” was launched in what year?

☐ Where is it believed that fireworks were invented?

☐ Which of these is found in the brain?

☐ Which of these is in North America?

☐ Which of Galileo’s achievements brought him into conflict with the church, resulting in his being confined to his house for the last years of his life?

☐ What is the closest planet to the sun?

☐ On which continent are the native fauna called ostrich, lion, giraffe and okapi?

☐ What do anthropologists study?

☐ In the Alfred Hitchcock film ”Psycho”, where did the murder take place?

☐ Where would one find a hypotenuse?

☐ Who has won the most Olympic Gold medals?

☐ What does the chemical symbol Fe stand for?

☐ Michael J Fox played which character in the ”Back to the Future” trilogy (1985-1990)?

☐ In medicine, what do the s BMI stand for?

☐ A single flame gas burner frequently used in student science laboratories is named after whom?

☐ A sabre is what type of weapon?

PART 3

Instructions

You now have the opportunity to earn up to $8 by writing a short essay to convince a group of reviewers of how well you performed on the test.

The reviewers are another group of QUT students participating in this experiment, in another room, as we speak. However, the reviewers have no experience in the test you just completed and we will not be providing the reviewers with any information about how you actually did in that test.

Your payment for this task will depend both on (A) how well the reviewers think you performed on the test and (B) on how convincing your essay is.

(A) If your score is rated in the 50% best scores of today's session participants by the reviewers, you will be paid $2 and an additional $2 if your score is rated in the 10% best scores.

(B) If your essay is rated in the 50% of the most convincing essays by the reviewers, you will get $2 and if your essay is rated in the 10% of most convincing essays, you will earn another $2, independent of how high they think you scored.

When your essay is done, please hit the "next" button.

Essay rules:

The content of your essay is not restricted in any way.

You are only forbidden to make threats, to reveal your identity, seat number or anything that might uncover your anonymity.

If you violate these restrictions you will not receive any payment at the end of the experiment.

Next

PART 4

Instructions

You now have the opportunity to earn additional money by estimating how many questions out of the 31 on the test you correctly answered. You will be paid as a function of the accuracy of your estimate.

You will earn $2 on this task if your estimate is in the 50% of most accurate estimates. You will earn another $2 if your estimate is in the 10% of most accurate estimates.

Please estimate how many questions you think you scored correctly on the general knowledge test :

4

Next

PART 5

Instructions

On the slider below, place yourself between 0% and 100% in terms of your performance relative to the other participants in this room.

For example, if you think you did better than 3/4 of your fellow participants but worse than 1/4 of them, you will choose 75%. Conversely, if you think that you did better than 1/4 of your fellow participants but worse than 3/4, you will choose 25%.

You will earn $2 on this task if the estimate of your rank compared to your fellow participants is in the 50% of most accurate estimates. You will earn an additional $2 if the estimate of your rank is in the 10% of most accurate estimates.

How do you think you did compare to the other students in this session in terms of number of items correct?

0

Next

PART 6

Instructions

Please rate your familiarity with each item by selecting the appropriate number from 0 to 4. If you are very familiar with an item, choose the number 4. If you have never heard of an item, choose the number 0.

Items	0 (Never heard of it)	1	2 (Somewhat familiar)	3	4 (Very familiar)
Houdini:	●	○	○	○	○
Charlotte Bronte:	○	●	○	○	○
Meta-toxin:	○	●	○	○	○
Myth:	○	○	●	○	○
Antigone:	○	●	○	○	○
Cholarine:	○	●	○	○	○
Alliteration:	○	●	○	○	○
Gail Brennan:	○	●	○	○	○
Queen Shattuck:	○	●	○	○	○
Lewis Caroll:	○	●	○	○	○
Free will:	○	○	●	○	○
Dale Carnegie:	○	●	○	○	○
Murphy's Last Ride:	○	○	●	○	○
Sentence stigma:	○	○	●	○	○
Bay of Pigs:	○	●	○	○	○

— Reviewers —

Instructions

Some QUT students are participating in an experiment in a different room, as we speak.

In this experiment, your fellow participants took a 31 question general knowledge test and we asked them to write an essay about their performance.

Before writing their essay, your they were allowed to check 10 questions and were told whether they answered each of these questions correctly or not.

Your task during this experiment will be to rate some of these short essays both on (A) how convincing you think the essay is and (B) how many questions you think the author answered correctly in the knowledge test.

Payment

Each guess of how many questions you think the author of the essay answered correctly in the knowledge test gives you an opportunity to earn extra money.

If your guess is exactly equal to the true performance of the essay writer or deviates from the true performance of the essay writer by 1 question (meaning that the true score if the essay writer was either 1 point above or below your guess), you will earn $10.

If your guess deviates from the true performance of the essay writer by 2 question (meaning that the true score if the essay writer was either 2 points above or below your guess), you will earn $8.

If your guess deviates from the true performance of the essay writer by 3 questions (meaning that the true score if the essay writer was either 2 points above or below your guess), you will earn $5.

If your guess deviates from the true performance of the essay writer by 4 or 5 questions (meaning that the true score if the essay writer was either 4 or 5 points above or below your guess), you will earn $2.

If your guess deviates from the true performance of the essay writer by more than 2 questions, you will not earn anything.

At the end of the experiment, the program will randomly select one guess and you will be paid according to that guess only. Therefore, you should treat every guess as if it is the guess that matters.

When you are ready to start, please hit the 'next' button.

Next

A.5 25-item version of the Over-Claiming Questionnaire

The OCQ scale measures the dispositional overconfidence. Participants are asked to rate how familiar they are with each item on a scale from 0 to 4 (0 being not familiar at all and 4 being very familiar), 8 of which are non-existent. The sum of ratings of nonexistent items constitutes the over-claiming index that we use as a measure of dispositional overconfidence. The Over-Claiming Questionnaire originally proposed by Paulhus et al. (2003) is composed of 150 items classified in 10 categories. In each categories, 3 out of the 15 items are non-existent. In this version, participants are asked to indicate how familiar they are with each item of the series on a 6-point Likert scale (Likert, 1932). The OCQ is placed before the general knowledge test in Study 1, whereas other control measures are collected at the end of the experiment. In our version, we added 'Australia' as an attention check.

Table A.11: OCQ items

Item	Foil or Real
Houdini	Real
Charlotte Bronte	Real
meta-toxins	Foil
Antigone	Real
cholarine	Foil
alliteration	Real
Gail Brennan	Foil
myth	Real
Queen Shattuck	Foil
Lewis Carroll	Real
free will	Real
Dale Carnegie	Real
Murphy's Last Ride	Foil
sentence stigma	Foil
Bay of Pigs	Real
hyperbole	Real
The Aeneid	Real
euphemism	Real
double entendre	Real
consumer apparatus	Foil
blank verse	Real
shunt-word	Foil
art-deco	Real
Artemis	Real
a cappella	Real

A.6 General Knowledge test items

The items that differ between the two studies are highlighted in bold. The correct answers are displayed in red.

Study 1

1. **Who was the first person to sign the American Declaration of Independence?**
   David Crockett / George Washington / John Hancock / Benjamin Franklin

2. The unit of electrical resistance was named after whom?
   Georg Simon Ohm / Benjamin Franklin / Guglielmo Marconi

3. Titan is a moon of which planet?
   Mars / Uranus / Saturn / Venus

4. How many pieces are on a chessboard at the start of a game?
   8 / 32 / 16 /64

5. Which of these is the largest in area?
   Spain / Texas / Tanzania / Afghanistan

6. **Who did George W. Bush beat for the US presidency in 2000?**
   Al Gore / John Kerry / John F. Kennedy / John McCain

7. At the opening ceremony of every Olympic Games when the athletes parade into the stadium, what is traditionally the first nation to enter?
   Australia / Simbabwe / Greece / Denmark

8. Which of these types of music did not originate in the Caribbean?
   Gregorian chant / Flamenco / Ska / Reggae

9. What is the name of the engraved stone, discovered in 1799, that provided a key to deciphering the languages of ancient Egypt?
   Babel Stone / Blarney Stone / Rosetta Stone / Talking Stone

10. **The portrait of which US statesman appears on the US $100 bill?**
    Abraham Lincoln / George Washington / Theodore Roosevelt / Benjamin Franklin

11. Which Scotsman took out a patent in 1876 that was the nucleus of the telephone?
    Alexander Fleming / Thomas Edison / George Stephenson / Alexander Bell

12. What is another name for a blood clot?
    Abrasion / Carcinoma / Bursitis / Thrombosis

13. Which of these countries is not landlocked? Paraguay / Bolivia / Andorra / Australia

14. The tibia and fibula are found where in the human body? Arm / Lower leg / Ribcage / Fingers

15. "Facebook" was launched in what year?
    1990 / 1994 / 2004 / 2009

16. Where is it believed that fireworks were invented?
    China / Mexico / Egypt / Greece

17. Which of these is found in the brain?
    Tibia / Thalamus / Vertebra / Humerus

18. Which of these is in North America?
    The Orzaks / The Urals / The Himalayas / The Pyrenees

19. Which of Galileo's achievements brought him into conflict with the church, resulting in his being confined to his house for the last years of his life?
    He attempted to measure the speed of light / He invented the thermometer / He said that Copernican view of the universe was correct / He attempted to measure the weight of air

20. What is the closest planet to the sun?
    Venus / Mercury / Saturn / Mars

21. On which continent are the native fauna called ostrich, lion, giraffe and okapi?
    Africa / South America / Australia / Asia

22. What do anthropologists study?
    Human Beings / Coal / Monkeys / Minerals

23. In the Alfred Hitchcock film "Psycho", where did the murder take place?
    In the bedroom / In the kitchen / On the front porch / In the shower

24. Where would one find a hypotenuse?
    Under the wing of a chicken / In the roof of a wooden building / As part of a right angled triangle / In a vehicle's gearbox

25. Who has won the most Olympic Gold medals? Larrisa Latynina, URSS / Paavo Nurmi, Finland / Michael Phelps, USA / Mark Spitz, USA

26. What does the chemical symbol Fe stand for?
    Iron / Gold / Silver / Cheese

27. Michael J Fox played which character in the "Back to the Future" trilogy (1985-1990)?
    Mickey McMouse / Marty McFly / Morris McAustin / Maurice McKee

28. In medicine, what do the s BMI stand for?
    Bionic Machine Implants / Biochemical Mortuary Investigators / British Medical Institute / Body Mass Index

29. A single flame gas burner frequently used in student science laboratories is named after whom?
    John Tilley / Michael Faraday / Sir Humphry Davy / Robert Bunsen

30. **A filibuster is typically found where?**
    A decision-making body / A hospital / A society party / A horse race

31. (*Attention Check*) What day comes after Tuesday?
    Monday / Wednesday / Thursday / Friday

Study 2

1. **What bird has the widest wingspan?**
   Albatros / Condor / Eagle / Vulture

2. The unit of electrical resistance was named after whom?
   Georg Simon Ohm / Benjamin Franklin / Guglielmo Marconi

3. Titan is a moon of which planet?
   Mars / Uranus / Saturn / Venus

4. How many pieces are on a chessboard at the start of a game?
   8 / 32 / 16 /64

5. Which of these is the largest in area?
   Spain / Texas / Tanzania / Afghanistan

6. **In Greek mythology, who was the multi-headed dog, encircled by a serpent, that guarded the portal to the underworld?**
   Minotaur / Rover / Cerebrus / Buccephalus

7. At the opening ceremony of every Olympic Games when the athletes parade into the stadium, what is traditionally the first nation to enter?
   Australia / Simbabwe / Greece / Denmark

8. Which of these types of music did not originate in the Caribbean?
   Gregorian chant / Flamenco / Ska / Reggae

9. What is the name of the engraved stone, discovered in 1799, that provided a key to deciphering the languages of ancient Egypt?
   Babel Stone / Blarney Stone / Rosetta Stone / Talking Stone

10. **The Dalai Lama is a high lama in which religion?**
    Buddhism / Toaism / Hinduism /Christianism

11. Which Scotsman took out a patent in 1876 that was the nucleus of the telephone?
    Alexander Fleming / Thomas Edison / George Stephenson / Alexander Bell

12. What is another name for a blood clot?
    Abrasion / Carcinoma / Bursitis / Thrombosis

13. Which of these countries is not landlocked? Paraguay / Bolivia / Andorra / Australia

14. The tibia and fibula are found where in the human body? Arm / Lower leg / Ribcage / Fingers

15. "Facebook" was launched in what year?
    1990 / 1994 / 2004 / 2009

16. Where is it believed that fireworks were invented?
    China / Mexico / Egypt / Greece

17. Which of these is found in the brain?
    Tibia / Thalamus / Vertebra / Humerus

18. Which of these is in North America?
    The Orzaks / The Urals / The Himalayas / The Pyrenees

19. Which of Galileo's achievements brought him into conflict with the church, resulting in his being confined to his house for the last years of his life?
    He attempted to measure the speed of light / He invented the thermometer / He said that Copernican view of the universe was correct / He attempted to measure the weight of air

20. What is the closest planet to the sun?
    Venus / Mercury / Saturn / Mars

21. On which continent are the native fauna called ostrich, lion, giraffe and okapi?
    Africa / South America / Australia / Asia

22. What do anthropologists study?
    Human Beings / Coal / Monkeys / Minerals

23. In the Alfred Hitchcock film "Psycho", where did the murder take place?
    In the bedroom / In the kitchen / On the front porch / In the shower

24. Where would one find a hypotenuse?
    Under the wing of a chicken / In the roof of a wooden building / As part of a right angled triangle / In a vehicle's gearbox

25. Who has won the most Olympic Gold medals? Larrisa Latynina, URSS / Paavo Nurmi, Finland / Michael Phelps, USA / Mark Spitz, USA

26. What does the chemical symbol Fe stand for?
    Iron / Gold / Silver / Cheese

27. Michael J Fox played which character in the "Back to the Future" trilogy (1985-1990)?
    Mickey McMouse / Marty McFly / Morris McAustin / Maurice McKee

28. In medicine, what do the s BMI stand for?
    Bionic Machine Implants / Biochemical Mortuary Investigators / British Medical Institute / Body Mass Index

29. A single flame gas burner frequently used in student science laboratories is named after whom?
    John Tilley / Michael Faraday / Sir Humphry Davy / Robert Bunsen

30. **A sabre is what type of weapon?**
    Rifle / Spear / Crossbow / Sword

31. (*Attention Check*) What day comes after Tuesday?
    Monday / Wednesday / Thursday / Friday

Appendix B

Appendix of Chapter 3

B.1 Summary Statistics

Table B.1 displays the average values for the number of correct answers in part II, the initial value of the group account, the proportion of females, the average age and average risk preferences by signals. We also show the percentage of female by signals. As expected, there is a significant difference in performance between participants who received a good signal and participants who received a bad signal (MW test: $p < 0.001$). In contrast, we do not find any significant differences in the value of the group account, percentage of female, age nor risk preferences.

Table B.1: Summary of the individual characteristics, by signals.

signals	N	mean perf. (part II)	group account (initial)	female (percentage)	Mean age	Mean risk
B	150	16.97***	24.37	54.67%	22.93	5.74
G	148	18.93	24.10	53.38%	23.09	6.17
all	298	17.93	24.23	54.03%	23.01	5.95

Note: Table B.1 displays the average values for the number of correct answers in Part II, the initial value of the group account, the proportion of females, age and risk preferences by signals. Stars indicate two-sample Mann-Whitney tests, comparing participants who received a good signal and participants who received a bad signal. * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

Table B.2 shows the same average values by combinations of signals. One-way between-subject ANOVAs show no significant difference between combinations of signals in terms of performance, initial value of the group account, percentage of female, age and risk preferences ((F(2, 146) = 0.20; $p = 0.815$; $p = 0.878$; $p = 0.919$; $p = 0.982$; $p = 0.134$, respectively).

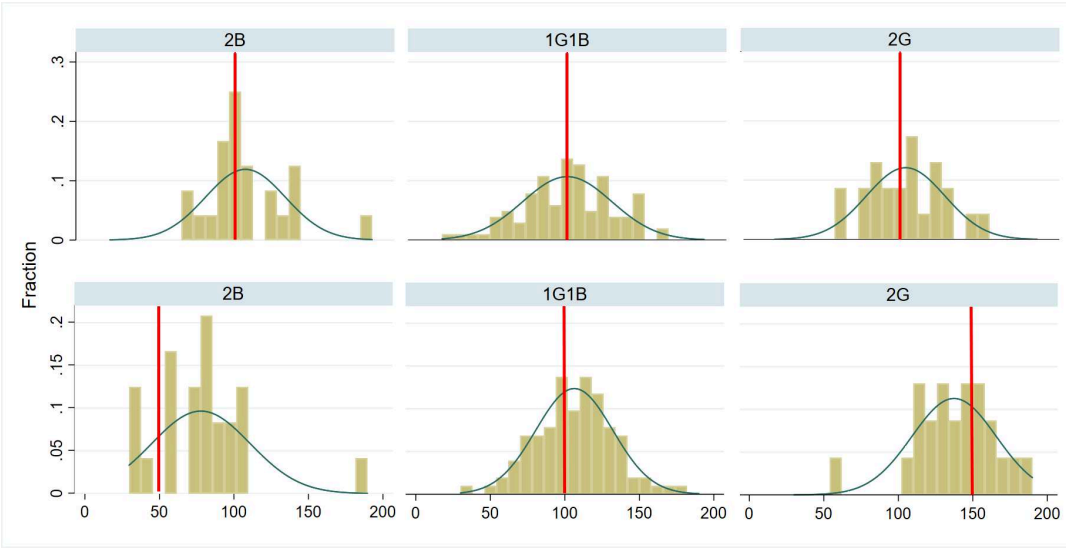
Table B.2: Summary of the individual characteristics, by combinations of signals.

Pairs of signals	nb. pairs	mean perf.	group account (initial)	female (percentage)	Mean age	Mean risk
BB	24	18.10	24.60	56.25%	22.23	5.88
		(0.594)	(0.862)	(0.072)	(0.418)	(0.261)
1G1B	102	18.09	24.26	54.41%	23.17	5.79
		(0.285)	(0.407)	(0.036)	(0.411)	(0.150)
GG	23	17.94	24.23	54.03%	23.01	5.95
		(0.235)	(0.342)	(0.029)	(0.317)	(0.119)

Note: Table B.2 displays the average values for the number of correct answers in Part II, the initial value of the group account, the proportion of females, age and risk preferences by signals.

B.2 Additional Analyses

B.2.1 Confidence and signals



Note: Vertical lines indicate the Bayesian theoretical average prior and posterior beliefs conditional on the combination of signals received.

Figure B.1: Distribution of the sum of prior beliefs (upper panel) and posterior beliefs (lower panel) within pairs, conditional on the signals received.

Figure B.1 shows the distribution of the sum of prior beliefs (upper panel) and posterior beliefs (lower panel) within pairs, conditional on the combination of signals received. The vertical lines indicate the Bayesian average prior and posterior beliefs. Before receiving the signals, participants' average belief should be roughly around 50% has they don't have any reason to believe that they performed better

or worse than their partner. Hence, the average sum of prior beliefs within pairs should be 100%. Participants are then given a binary signal that gives them information about their relative performance with a 75% accuracy. When receiving a bad signal, a Bayesian participant should update his beliefs from 50% to 25%. In contrast, a Bayesian participant should update his beliefs from 50% to 75% when receiving a good signal. Hence, the average theoretical sum of posterior beliefs should be 50% for participants who received two bad signals, 100% for participants who received two opposite signals and 150% for participants who received two good signals.

Table B.3 displays the mean values and standard errors of the average sum of posterior beliefs of participants i and j from the pair $\{i, j\}$ by combinations of signals. Results from two-sample Mann-Whitney tests show that pairs of participants who received two bad signals hold significantly lower beliefs than pairs who received one good and one bad signal ($p < 0.001$). In contrast, pairs of participants who received two good signals hold significantly higher beliefs than pairs who received one good and one bad signal ($p < 0.001$).

Table B.3: Average sum of posterior beliefs within pair, conditional on the signals received.

	2B	1B1G	2G
$Belief_i + Belief_j$	78.25***	103.35	137.43***
(posterior)	(4.720)	(1.803)	(4.128)
N.	24	102	23

Note: Table B.3 summarizes posterior beliefs at the pair level for pairs of participants who received two good signals (2G), one good and one bad signal (1G1B) and two bad signals (2B). Standard errors in parentheses. Stars indicates the results of two-sample Mann-Whitney tests between pairs with one good and one bad signals and pairs with another combination of signals. *** p <0.01, ** p <0.05, * p <0.10.

B.2.2 Confidence, agreements and delays

Figure B.2 shows Kaplan-Meier survival estimates of the probability that a pair of participants survive beyond a certain stage of the negotiation process ("survival function"). In other words, each curve represents the likelihood for a specific combination of signals that an agreement is reached in a given stage of the negotiation process.

We can see that pairs of participants who received two good signals are more likely to survive until the end of the negotiation process (i.e. less likely to reach

an agreement) than pairs of participants who received one good and one bad signals (log-rank test: $p = 0.040$). We find no significant differences between participants who received two bad signals and pairs of participants who received any other combination of signals (log-rank tests: 2B *vs.* 1G1B, $p = 0.988$; 2B *vs.* 2G, $p = 0.131$). The absence of significant differences between the survival function of pairs who received two good signals and the survival function of pairs who received two bad signals is likely driven by the low number of observations compared to pairs who received one good and one bad signal. Indeed, the p-value of the log-rank test is close to conventional significance levels.

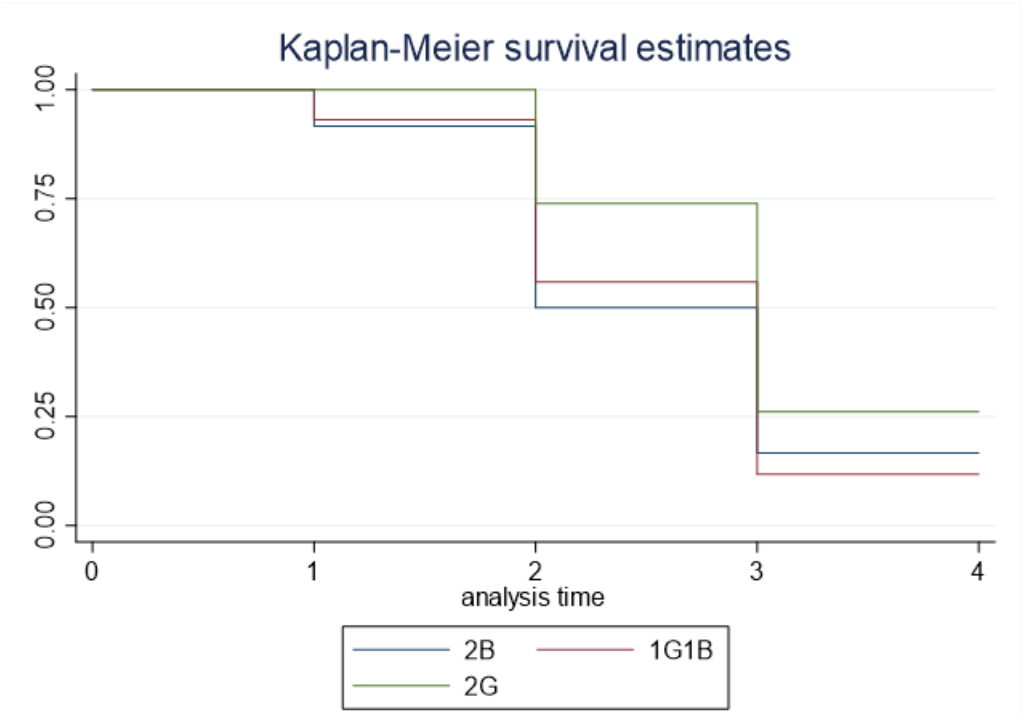


Figure B.2: Survival function of the rate of agreements for pairs of participants who received two bad signals (in blue), pairs who received one good and one bad signal (in red) and pairs who received two good signals (in green).

Table B.4 reports the same analyses as in Table 3.2 using the combinations of signals at the pair levels instead of the sum of participants beliefs in the same pair. Consistent with results from Table 3.2, Table B.4 shows that pairs of participants who received two good signals are marginally less likely to reach an agreement.

Table B.4: Effect of a combination of signals on delays and impasses.

Dep. var:	**Rate of Agreements** (t=stage)
1G1B	*Ref.*
2G	-0.050*
	(0.258)
2B	-0.506
	(0.273)
Obs.	149

Note: Table B.4 reports the estimates of Cox regressions with proportional hazards of each combination of signals on the rate of agreements. The unit of observation is one pair. Standard errors in parentheses. *** p <0.01, ** p <0.05, * p <0.10.

B.2.3 Confidence and Individual Outcome

Table B.5 displays the percentage of participants who switched from the high share to the low share, conditional on the signal they received. Column (1) displays the proportion for participants who agreed in stage 2. Column (2) displays the proportion for participants who agreed in stage 3. Column (3) displays the proportion for participants who agreed in stage 2 or in stage 3. Column (4) displays the proportion for all participants who did not agree in stage 1.

Table B.5: Proportion of participants who switched to the low share across stages, by signals.

Signal	Switch to low share in			
	Stage 2	Stage 3	stage 2 or 3	All
	(1)	(2)	(3)	(4)
B	68.97%***	49.18%	58.82%***	50.36%***
	(0.061)	(0.064)	(0.045)	(0.042)
G	28%	49.25%	40.17%	33.33%
	(0.063)	(0.061)	(0.045)	(0.040)
Obs.	108	128	119	139

Note: Table B.5 shows the proportion of participants who switched from the high share to the low share. Column (1) displays the proportion for participants who agreed in stage 2. Column (2) displays the proportion for participants who agreed in stage 3. Column (3) displays the proportion for participants who agreed in stage 2 or in stage 3. Column (4) displays the proportion for all participants who did not agree in stage 1. Stars indicate the results of two-sample tests of proportion between G_iB_j and B_iG_j. *** p <0.01, ** p <0.05, * p <0.10.

We find that participants who received a good signal are significantly less likely to switch in stage 2 than participants who received a bad signal (PR test: $p < 0.001$). Interestingly, this difference disappears in stage 3 (PR test: $p = 0.993$). Considering stage 2 and 3 together, participants who received a good signal are less likely to switch than participants who received a bad signal, both when considering participants who did not reach an agreement and participants who reached an agreement in either stage 2 or 3 only (PR tests: $p = 0.004$ in both cases).

Table B.6: Effect of relative beliefs on participants' payoff from the negotiation.

Dep. var: fraction of group account	All		Agreements only	
	(1)	(2)	(3)	(4)
B_iG_j	*Ref.*	*Ref.*	*Ref.*	*Ref.*
B_iB_j	0.032	0.031	0.062**	0.058**
	(0.043)	(0.044)	(0.024)	(0.025)
G_iB_j	0.098***	0.097***	0.111***	0.108***
	(0.033)	(0.033)	(0.038)	(0.037)
G_iG_j	-0.092*	-0.095*	-0.046	-0.053
	(0.051)	(0.052)	(0.048)	(0.050)
Age		-0.001		-0.003
		(0.002)		(0.002)
Female		-0.020		-0.003
		(0.028)		(0.027)
Risk preferences		0.002		0.009
		(0.007)		(0.007)
Constant	0.354***	0.385***	0.401***	0.412***
	(0.021)	(0.071)	(0.019)	(0.066)
Obs.	298	298	254	254

Note: Table B.6 shows the results of the OLS estimations of the percentage of the group account received at the end of the negotiation on the different combinations of signals. Standard errors in parentheses. *** p <0.01, ** p <0.05, * p <0.10.

Table B.6 replicates the results of Table 3.6, using the different combinations of signals as independent variables instead of the differences in beliefs between participants from the same pair. The independent variables include dummies of each combination of signals. All regressions are clustered at the pair level. Models (1) and (2) show the results for all pairs. Models (3) and (4) shows the results for participants who reached an agreement only. In models (2) and (4), we control

for participants demographics (sex, age and risk preferences).

Table B.6 shows that participants who received a good signal are more likely to end up with a larger percentage of the initial group account than their partners for pairs of negotiators who received opposite signals and the results are significant at the 1% level. In addition, we find that participants who received a good signal when their partner also received a good signal are significantly worse off than any other combination (Chi2 tests: G_iG_j vs. B_iB_j, $p = 0.040$; G_iG_j vs. G_iB_j, $p < 0.001$; G_iG_j vs. B_iG_j, $p = 0.074$) and those results holds when controlling for participants demographics.[1] Interestingly, while participants who received a good signal when their partner also received a good signal are not significantly worth off than participants who received a bad signal when their participants received a good signal when considering only pairs of participants who reach an agreement, they are still significantly worse off than participants who received a good signal when their partner received a good signal and than participants who received a bad signal when their partners' also received a bad signal (Chi2 tests: G_iG_j vs. B_iB_j, $p = 0.022$; G_iG_j vs. G_iB_j, $p = 0.002$; G_iG_j vs. B_iG_j, $p = 0.339$) and those results holds when controlling for participants demographics.[2]

B.3 Biases in Beliefs Updating

Even though economic models assume that people update their beliefs according to Bayes rules, data show that it is often not the case. Empirical evidence show that people tend to update conservatively, by responding too little to new information.[3] In addition, other experiments have shown that participants update positive and negative feedback asymmetrically (Eil and Rao, 2011; Mobius et al., 2011; Sharot et al., 2012; Wiswall and Zafar, 2015). Eil and Rao (2011) and Mobius et al. (2011) investigate beliefs updating in a context where participants receive self-relevant feedback. In their experiments, participants are ranked based on an IQ test and are asked to state their beliefs about their performance compared to the other participants. The authors found that participants give higher weight to good feedback, relative to bad feedback. However, the results on asymmetric updating in the literature are mixed. Schwardman and van der Weele (2019), for instance, also use an IQ-relevant task but do not find evidence of asymmetric updating in their data.

[1] (Chi2 tests: G_iG_j vs. B_iB_j, $p = 0.038$; G_iG_j vs. G_iB_j, $p < 0.001$; G_iG_j vs. B_iG_j, $p = 0.071$).
[2] (Chi2 tests: G_iG_j vs. B_iB_j, $p = 0.021$; G_iG_j vs. G_iB_j, $p = 0.002$; G_iG_j vs. B_iG_j, $p = 0.283$).
[3] Recent examples include: Eliaz and Schotter (2010); Ertac (2011); Mobius et al. (2011); Buser et al. (2016); Coutts (2019); Hoffman (2016); Schwardman and van der Weele (2019); Ambuehl and Li (2018).

In this section, we investigate whether conservatism (i.e. participants place too little weight on new information) and asymmetric updating (participants place more weight on good signals compared to bad signals) exist in our sample. We follow Mobius et al. (2011) in running logit regressions, which are based on a linearized version of Bayes' formula. The model is given by:

$$logit(\mu_{i,post}) = \delta logit(\mu_{i,prior}) + \beta_G I(s_i = G)\lambda_G + \beta_B I(s_i = B)\lambda_B + \epsilon_i \quad \text{(B.1)}$$

with $logit(x)$ =ln$(x/(1-x))$. In our design, $\mu_{i,prior}$ represent the prior belief (i.e. before the signal is observed) of participant i regarding the probability that his performance in part II was higher than his partner performance in part II. $\mu_{i,post}$ is participant i's posterior belief (i.e. after the signal is observed).$\lambda_G = \lambda_B$ is the log of the likelihood ratio (3 in our case). $I(si = G)$ is an indicator variable that equals 1 if participant i received a good signal and 0 otherwise, and $I(si = B)$ is an indicator variable that equals 1 if participant i received a bad signal and 0 otherwise. If participants are perfect Bayesians, we should observe $\delta, \beta_G, \beta_B = 1$. If participants exhibit conservatism, we should observe both $\beta_G < 1$ and $\beta_B < 1$. If participants exhibit asymmetric updating, we should observe $\beta_G > \beta_B$.

Table B.7: Belief Updating

	Posterior beliefs	
	(1)	(2)
δ	0.812***	0.849***
	(0.049)	(0.049)
β_G	0.282***	0.290***
	(0.020)	(0.019)
β_B	0.220***	0.233***
	(0.020)	(0.019)
Obs.	283	275

Note: Model (1) displays the estimates for the entire sample, whereas model (2) excludes participants who update in the wrong direction. H0: coefficient equals 1. Standard errors in parentheses. * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

Table B.7 shows the results of OLS regressions, where stars indicate rejections of the null hypothesis that the coefficient is 1 at different levels of confidence. Column (1) displays the estimates for the entire sample, whereas column (2) excludes participants who update in the wrong direction.

First of all, results from Table B.7 show that we reject the null hypothesis that $\delta = \beta_G = \beta_B = 1$. These results suggest that participants are conservative on average. We then test our hypothesis that $\beta_G > \beta_B$. We found that this is the case (Chi2 tests: $p = 0.026$ and $p = 0.040$, respectively), confirming our hypothesis that our participants exhibit asymmetric updating on average. There results are in line with the literature.

B.4 Messages Content

Participants were allowed to write a message to explain to their partner why they choose a particular share. Since most participants claimed the largest share in stage 0, this section provide a summary of the main reasons provided by participants to justify their choice of the largest share. The messages were coded using dummy variables equals to 1 if the following categories were mentioned: Blackball strategy (the participant explicitly states that he will not back down from her initial claim), Merit (the participant explicitly states that he deserves the largest share based on his performance at the task), Improvement (the participant mentions the fact that his performance improves from the first part to the second), Outside lab reasons (the participant uses arguments from his personal life or background to justify his choice) and Risk (the participant mentions the risk to loose everything if they fail to reach an agreement). One message can belong to several categories. Messages that did not match any of these categories were classified as "Other".

We first investigate whether there are significant differences in the essay content of participants depending on the signal they received. Figure B.3 shows that participants who received a good signal go for the blackball strategy (marginally) more often than participants who received a bad signal and mention more often that their performance justifies their choice, as well as the risk of ending up empty-handed if they fail to reach an agreement (two-sample tests of proportion: p=0.060, p=0.006 and p <0.001, respectively). In addition, messages of participants who received a good signal are significantly longer on average than messages of participants who received a bad signal (MW test: 30.31 words vs. 42.76 words; p=0.035).[4] Correlation analyses between participants' score during the second part of the experiment and the various messages categories reveals that participants with higher score talk about merit more and refer less to outside-

[4]Participants who failed to reach an agreement in the first stage of the negotiation process were allowed to communicate with their partner via an interaction chat box. However, the analysis of the chat content does not provide any additional findings.

the-lab arguments.

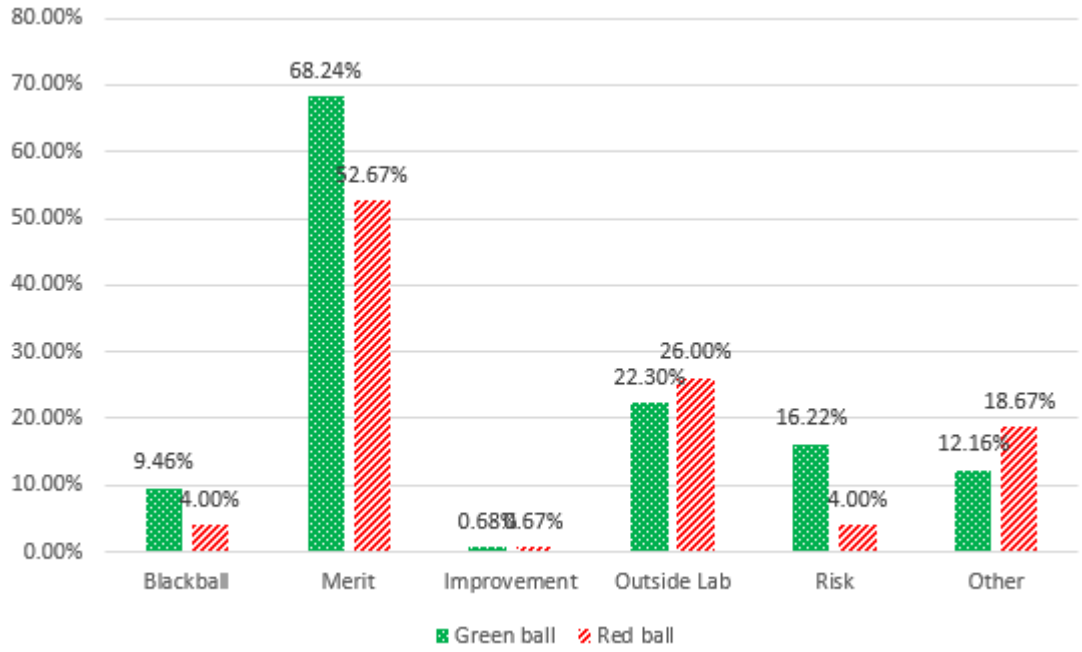


Figure B.3: Arguments mentioned in the messages to justify the choice of the largest share.

We also investigate the relationship between the chat variables and our main variables of interest. These exploratory analyses reveal that using a blackball strategy is negatively correlated with the likelihood to reach an agreement (Spearman correlation: r_s=0.15; p =0.008) but positively correlated with the likelihood to receive the high share of the group account, conditional on reaching an agreement (SC: r_s=0.20; p =0.002). We do not find any significant relationship between the amount received from the negotiation process and the messages variables.

To summarize, participants' performance and confidence affect the content of their messages and, henceforth, their attempt to persuade others. While more competent participants focus on their performance as their main arguments, less competent participants use more outside-the-lab arguments. However, we do not find evidence of a relationship between these categories and the outcome of the negotiation process.

B.5 Instructions

INTRODUCTION[5]

Welcome to this experiment on decision making.

Please turn off your phone and put it away.

Please do not communicate with the other participants in this session except if you are explicitly told to do so in the instructions.

During this session, you can earn money.
Your payoffs will depend on your decisions and the decisions of the other participants in this session.

All your decisions are anonymous.

You will receive a show up fee of 5 euros for being on time at the experiment.

At the end of the experiment, you will be paid privately in cash in a separate room.

This session is composed of four parts.
You will receive the instructions for part 1 at the end of this introduction.
You will receive the instructions for each following part after you finish the previous part.
The instructions for Part 1, 2 and 4 will be given to you on papers, and instructions for Part 3 will be directly displayed to you on your computer screen.

If you have questions during this experiment, you can raise your hand or press the red button on your left and the experimenter will answer you in private.

INSTRUCTION PART 1

For this part of the experiment, you will perform a task individually. By working on the task, you will earn points. At the end of the experiment, each point will be converted according to the following exchange rate: 1 point = 0.20€

[5]Note that this is an English translation as the original experiment was conducted in French.

The task

Your task is to answer 10 general knowledge questions in the form of a Multiple-Choice Questionnaire. These questions are the same for every participant.

You will earn 1 point for each correct answer.
You will have 15 seconds per questions to make your decision. Once you have made your decision, press the 'next' button to start the next question.

If you selected an answer but failed to press the "next" button, the computer program will record your answer.

If you fail to select an answer before the end of the 15 seconds, the next question will start automatically, and you will not receive any points for that question.

* * *

If you have any questions, please raise your hand and the experimenter will answer you in private. When everyone is ready, Part 1 will start.

INSTRUCTION PART 2

Group formation

For this part of the experiment, you will be matched with another participant according to your performance in the previous knowledge test in Part 1. The participant who received the best score in Part 1 will be matched with the participant who received the second-best score in Part 1. The participant who received the third best score in Part 1 will be matched with the participant who received the fourth best score in Part 1 and so on until all the participants in the session are matched with another participant. Thus, you will be matched with a participant whose performance in Part 1 is ranked as close as possible to yours.

Task

In this Part, you and your partner will undertake the same type of general knowledge test as in Part 1. The MCQ test is composed of 30 questions. As in Part 1, these questions are the same for every participant. You will first undertake the task individually. However, your earnings for this Part will depend both on your performance and your partner's performance in this part. A group account will be allocated to each pair of participants. Each of your correct answers will earn 1 point to your group account. Similarly, each of your partner's correct answer will also earn 1 point to your group account.

At the end of the task, the total points you and your partner have earned will be multiplied by a random number between 0.85 to 1.15, which is drawn by the computer, and then converted to Euro according to following exchange rate: 1 point = 0.67€.

Group Account
The final value of your group account is determined by:

- The number of correct answers you provide in Part 2.
- The number of correct answers your partner provides in Part 2.
- The random number/multiplier the computer drew for your group.
- The exchange rate (1 point = 0.67 euro)

The random multiplier for your group is determined as follows: for each pair of participants, the computer program will randomly draw a ball from an urn containing 31 balls numbered from 0.85 to 1.15. The number on your ball will define the random multiplier for your group. For example, if the computer program draws a ball labelled 1.05, the value of your group account will be equal to the total points collected by you and your partner, times 1.05. You will not know which numbered ball was drawn by the computer program.

In summary, the final value of your group account is:
(the number of correct answers you received in Part 2 + the number of correct answers your partner received in Part 2) * the number drawn by the computer for your group * 0.67

At the end of this experiment, you and your partner will have to decide how to split the money in this group account. You will receive more information about the split of the group account at the beginning of Part 4.

To summarize:

- You will be paired with a participant whose performance in Part 1 as close as possible to yours.
- Each of your correct answers will earn 1 point to your group account.
- Each of your partner's correct answers will earn 1 point to your group account.

- The total points you and your partner have earned will be multiplied by a random number between 0.85 to 1.15, and then converted to Euro with an exchange rate of 1 point = 0.67 euro.
- At the end of the experiment, you will have to decide how to allocate the group account between the two of you.

* * *

If you have any questions, please raise your hand and the experimenter will answer you in private. When everyone is ready, Part 2 will start.

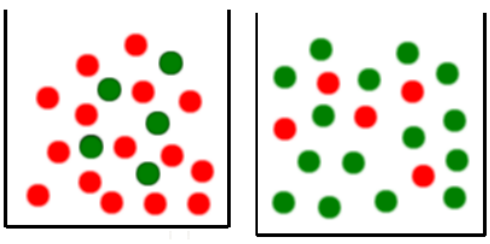
INSTRUCTION PART 3 [on screen]

[belief 1]
You finished part 2. Before we continue, we would like you to estimate how many questions you think you answered correctly. Your estimation will be rewarded as follow:

- If your estimation is exact or deviates from your performance by only 1 answer (i.e. your estimation deviates from your actual performance by more or less one correct answer), you will receive 1 euro.
- If your estimation deviates from your performance by 2 answers (i.e. your estimation deviates from your actual performance by more or less 2 correct answers), you will receive 0,50 euro.
- If your estimation deviates from your performance by more than 2 answers, you will not receive or lose anything.

[belief 2]
On the slider below, place the percentage of chance that you performed better than your partner between 1% and 100%.

[belief 3]
In this part of the experiment, you will see a ball drawn from one of the following urns:



- If you provided fewer correct answers that your partner, your ball will be drawn from an urn containing 15 red balls and 5 green balls. Thus, you will be more likely to see a red ball.
- If you provided more correct answers that your partner, your ball will be drawn from an urn containing 5 red balls and 15 green balls. Thus, you will be more likely to see a green ball.
- If you provided the exact same number of correct answers than your partner, your ball will be drawn from one of these two urns with a 50/50 chance; You will see a green or a red ball with the same probability.

Hence, a green ball means that you are likely to have provided more correct answers than your partner and a red ball means that you are likely to have provided fewer correct answers than your partner. Press the "next" button when you are ready to see your ball.

The computer program randomly drew a ball. The ball is red (green), which means that you are more likely to have provided fewer (more) correct answers than your partner.

Before we continue, we would like to estimate again the likelihood that your performance was better than your partner's.

Indicate the percentage chance that you provided more correct answers than your partner.

INSTRUCTION PART 4

In this part of the experiment, you and your partner have to split the group account. The value of your group account is displayed on your screen. **You can only split the group account according to the following allocation: one of you will receive 70% (share A) of the group account; the other will receive 30% (share B) of the group account.**

First, you and your partner will have to claim a share (either A or B). Together with this initial claim, you both can send a message to your partner explaining why you think you should have the share you claimed:

- If you and your partner choose different shares, you will both receive the share you claimed. Part 4 will be over.

- If you and your partner both choose the same share, you will enter into the Negotiation Stage.

NEGOTIATION PHASE

In this stage, you and your partner will have 3 minutes to agree on the split of the group account. During these 3 minutes, you will be able to negotiate with your partner via a chat box.

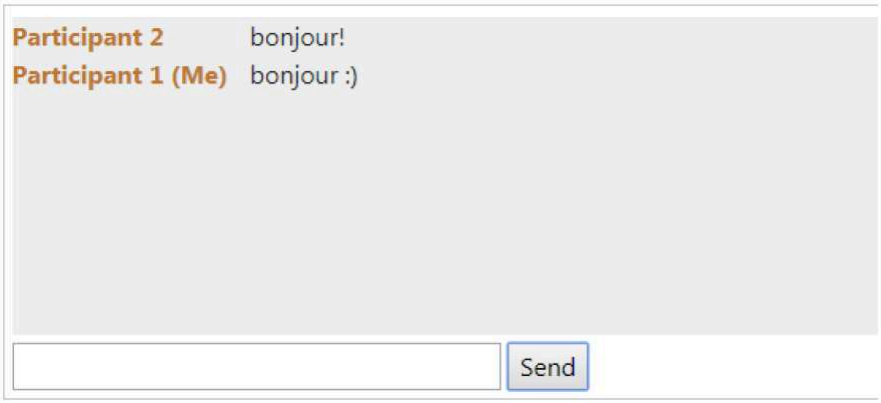
Interface:
You can see below an example of the decision screen you will see during the negotiation Stage:

In this example, the value of the group account is 20 euros. At the top of your screen, you can see the shares labeled A and B and their respective value. Share A corresponds to 70% of the group account (14 euros in this example) and share B corresponds to 30% of the group account (6 euros in this example).

You can use button A and button B in the middle of your screen to claim either share A or share B. When you click on either A or B, your choice will appear under 'your choice' in the table. In the example below, you chose A. Therefore, A appears under 'Your choice'.
Similarly, when your partner clicks on either A or B, her/his choice will appear under 'The other participant's choice' in the table. In the example below, your partner chose A. Therefore, A appears under 'The other participant's choice'.

By default, the choice you made at the beginning of this part will appear under 'your choice' on your screen and the choice your partner made at the beginning of this par will appear under 'the other participant's choice' on your screen. This information will also appear on your partner's screen.

You can use the chat box to negotiate with your partner. Here is a screenshot of the interactive chat box:



After you type your message in the text field, you can send it to your partner by pressing the "send" button.

Consequence:

- If either you or your partner switch and choose **a different share before the end of the 3 minutes**, you will both receive the share that you claimed, and this Part will be over.
- If you and your partner **stick to the same share at the end of 3 minutes**, you will be given 30 additional seconds to try to reach an agreement. **HOWEVER, for each second that passes, both shares will decrease proportionally,** and you will no longer be able to communicate with your partner.

- If you and your partner reach an agreement before the clock reaches 0, the counter will stop. The one who chose A will earn the remaining money allocated to share A and the other one will earn the remaining money allocated to share B.
- You can change your mind at any time by clicking the buttons A or B at the bottom of your screen.
- If the clock reaches 0 before you reach an agreement, you will both end up with 0 euro.

Chat and Messages:
The messages and chat rules are the following:

- You are not allowed to discuss the color of the ball you received in the previous part of the experiment or to give hints regarding the color of the ball.
- You are forbidden to make threats, to reveal your identity, seat number or anything that might uncover your anonymity.

If you violate these restrictions, you will not receive any payment you made during this part of the experiment.

In summary:

- You need to indicate what share of the group account you wish to receive and send a message to your partner explaining your choice.
- If you and your partner choose differently, you will both receive the share that you claimed.
- If you and your partner both choose the same share, you will have 3 minutes to chat, negotiate and agree on the split of the group account.
- If you and your partner still cannot reach an agreement at the end of 3 minutes, you will be given 30 additional seconds to try to reach an agreement.
- However, during these 30 seconds, both shares will decrease proportionally, and you will no longer be able to communicate with your partner.

* * *

Please read these instructions carefully again. If you have any questions, please raise your hand and the experimenter will answer you in private. Part 4 will start when all the questions are answered.

B.6 General Knowledge Test Items

This section presents the questions of the general knowledge tests for part I and part II of the experiment.[6] Under each question, we report the four options that were shown to the participants. The correct answer is displayed in red.

Questions General Knowledge Test - Part 1:

1. What is the name for the process by which heat is transferred by the motion of a fluid?
   conduction / convection / raditation / dissipation

2. At the opening ceremony of every Olympic Games when the athletes parade into the stadium, what is traditionally the first nation to enter?
   France / Zimbabwe / Greece / Denmark

3. What do anthropologists study?
   human beings / coal / monkeys / minerals

4. Who, in 1831, first demonstrated that the motion of a conductor in a magnetic field generates an electric current?
   Isaac Newton / Humphrey Davy / Ernest Rutherford / Michael Faraday

5. What is the name of the engraved stone, discovered in 1799, that provided a key to deciphering the languages of ancient Egypt?
   Babel stone / Blarney stone / Rosetta stone / talking stone

6. What French military unit was established in 1831 to enable people from other countries to serve in the French Armed Forces, commanded by French officers?
   the foreign army / the foreign legion / the foreign squad / the foreign forces

7. Which of Galileo's achievements brought him into conflict with the church, resulting in his being confined to his house for the last years of his life?
   He attempted to measure the speed of light / He invented the thermometer / He said that Copernican view of the universe was correct / He attempted to measure the weight of air

8. Which of these musical terms means the loudest?
   Mezzo forte/ Mezzo piano / Forte / Piano

9. In the Alfred Hitchcock film "Psycho", where did the murder take place?
   in the bedroom / in the kitchen / in the entrance / in the shower

[6] Note that this is an English translation as the original experiment was conducted in French.

10. In chess, what piece is allowed to jump over other pieces?
    the bishop / the knight / the rook / the pawn

Questions General Knowledge Test - Part 2:

1. John Milton created what name for the capital of Hell in his poem "Paradise Lost"?
   Dystopia / Bedlam / Chaos / Pandemonium

2. Tolstoy's book "War and Peace" is set when?
   100 Year's War / World War I / Crimean War / Napoleonic Wars

3. What is the name for the region of an astronomical object from which externally received light originates, which extends into a star's surface until the gas becomes opaque?
   chromosphere / corona / photosphere / cretaceous

4. Which South American country extends the furthest east?
   Argentina / Brazil / Uruguay / Bolivia

5. Which of these territories has the northernmost capital city?
   Iceland / Sweden / Russia / Canada

6. What are formed from linear chains of amino acids?
   Proteins / Carbohydrates / Red blood cells / Vitamins

7. Which of these characters can be found in Stendhal's book "The Red and The Black"?
   Julien Sorel / Pierre Rougon / Charles Bovary / Meursault

8. The unit of electrical resistance was named after whom?
   Alessandro Volta / Simon Ohm / Benjamin Franklin / Guglielmo Marconi

9. Titan is a moon of which planet?
   March / Uranus / Saturn / Venus

10. How many pieces are on a chessboard at the start of a game?
    8 / 32 / 16 / 64

11. Which of these is the largest in area?
    Spain / Texas / Algeria / Afghanistan

12. Which of these types of music did not originate in the Caribbean?
    Zouk / Flamenco / Ska / Reggae

13. Which Scotsman took out a patent in 1876 that was the nucleus of the telephone?
    Alexander Fleming / Thomas Edison / George Stephenson / Alexander Bell

14. "Facebook" was launched in what year?
    1994 / 2004 / 2009 / 1999

15. Which revered world figure celebrated his 95th birthday, in hospital, in July 2013?
    The Dalai Lama / Nelson Mandela / Ban Ki Moon / Pope Francis I

16. What is the closest planet to the sun?
    Venus / Mercury / Saturn / March

17. Who has won the most Olympic Gold medals?
    Paavo Nurmi, Finland / Michael Phelps, USA / Larissa Latynina, URSS / Mark Spitz, USA

18. In medicine, what do the initials BMI[7] stand for?
    Implants Mécaniques Corporels / Investissment Micro-Chimiques / Institut Médicale Canadien / Indice de Masse Corporelle

19. A single flame gas burner frequently used in student science laboratories is named after whom?
    John Tilley / Michael Faraday / Sir Humphry Davy / Robert Bunsen

20. Who played "Charlie" in the 2005 film "Charlie and the Chocolate Factory"?
    Johnny Depp / Macauley Culkin / Freddie Highmore / David Kelly

21. Conventionally, Lent, the period of the Christian calendar leading up to Easter is how long?
    One week / one month / 40 days / 15 days

22. Who was the Roman god of wine and fertility?
    March / Jupiter / Bacchus / Quirinus

23. Where is it believed that fireworks were invented?
    China / Mexico / Egypt / Greece

24. Which of these is found in the brain?
    Cuboid / thalamus / fibula / humerus

25. Which of these is in North America?
    The Ozarks / the Ural / the Himalayas / the Pyrenees

[7] IMB in French.

26. What science features in the ”Indiana Jones” film series?
    Physics / archaeology / physiotherapy / astronomy

27. What does the chemical symbol Fe stand for?
    Iron / gold / silver / charcoal

28. What does the ”B” stand for in the acronym ”FBI”?
    Bureau / Baltimore / Business / Bluster

29. Seth MacFarlane is the creator of which of these TV series?
    Beavis and Butthead / the Simpsons / South Park / the Griffins

30. The Richter scale measures the intensity of what?
    Rain / wind / earthquakes / tornados

Appendix C

Appendix of Chapter 4

C.1 Subject Pool

Table C.1 displays the number of participants, the number of groups, the average score, the proportion of female, the average age, the average risk aversion and the average social dominance for both treatments by session size. We don't find significant differences in the distribution of participants in terms of performance between treatments (two-sample MW test: $p = 0.182$). A two-sample test of proportions showed that there is no significant difference in the proportion of male and female between treatments ($p = 0.745$). There are no significant difference in the distribution of participants in terms of age between treatments (MW test: $p = 0.240$). Finally, we find no significant difference in the distribution of participants in terms of risk preferences and pre-disposition for social dominance between treatments (MW test: $p = 230$ and $p = 0.446$, respectively).

Table C.1: Summary of individual characteristics, by treatments.

	session size	Number of individuals	Number of groups	Mean score	Females (Percentage)	Mean age	Mean risk	Mean AMS
SI	16	32	8	11.31	62.5%	25.16	6.28	24.13
	12	60	15	11.85	56.67%	22.92	5.63	24.15
	8	24	6	13.29	41.67%	24.58	6.88	24.88
	All	116	29	12	55.17%	23.88	6.07	24.29
AI	16	48	12	11.56	60.42%	24.75	6.29	24.13
	12	36	9	11.69	44.44%	24.58	6.44	24
	8	40	10	11.70	65%	22.95	6.45	25.45
	All	124	31	11.65	57.26%	24.12	6.39	24.52

Note: Table C.1 displays the number of particiapnts, the number of groups, the average score, the proportion of female, the average age, the average risk aversion and the average social dominance for both treatments by session size.

C.2 Additional Analyses

C.2.1 Summary Statistics

Figure C.1 display the distribution of performances (white bars) and belief about these performances (dark bars) for the SI treatment (left panel) and the AI treatment (right panel). We can see that the distribution of beliefs is shifted to the right compared to the distribution of performance and this difference is significant in both treatments (Wilcoxon matched-pairs signed-ranks tests: $p < 0.001$ for both treatments), supporting results from section 4.3.1 that participants are overconfident on average in both treatments.

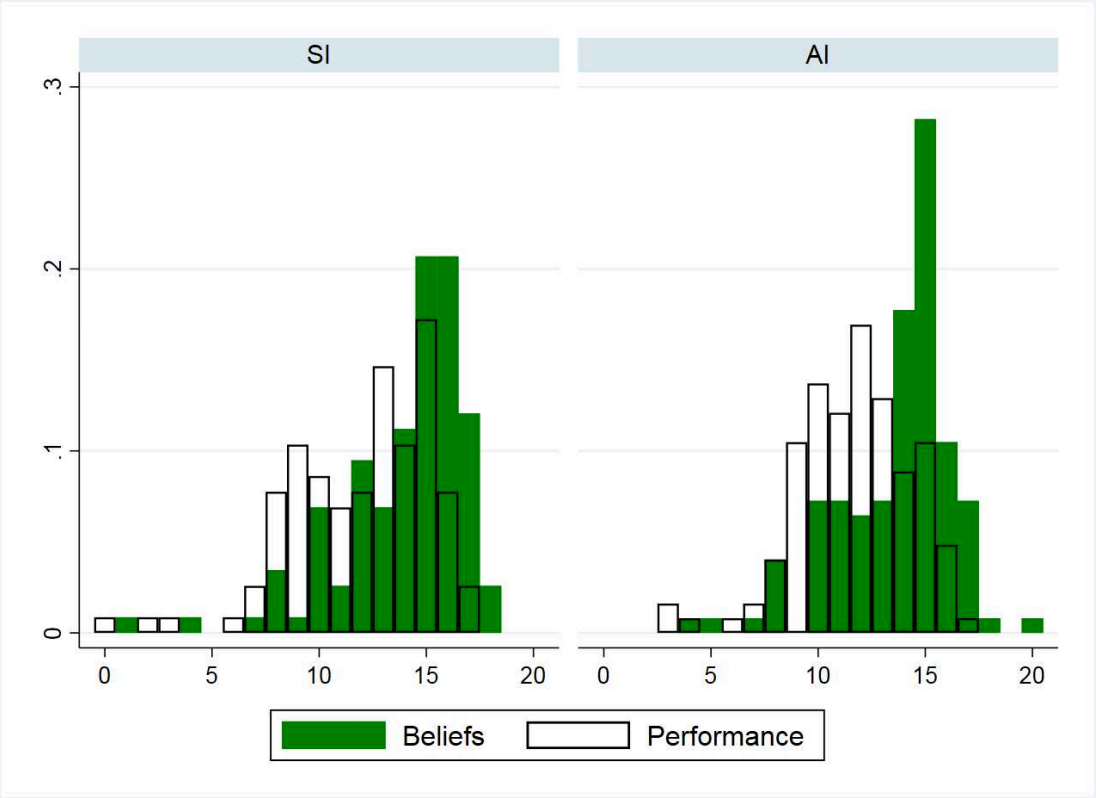


Figure C.1: Distributions of performance and beliefs about performance in both treatments.

Table C.2 displays the mean values (with standard errors in parentheses) of our variables of interest across both treatments. Performance (1) is measured as the number of matrices correctly solved by the participants. We elicit participants' beliefs about their performance (2) and the percentage of participants they outperformed (3). We also elicit participants' beliefs about the performance of the other participants in the session (4).

Table C.2: Mean values and standard errors of main variables.

	SI	AI	Δ	p-values
Performance (1)	12	11.65	0.355	0.353
	(0.300)	(0.239)	(0.381)	
(leaders only)	14.17	12.23	-1.94	0.001
	(0.391)	(0.422)	(0.577)	
Belief about performance (2)	13.97	13.63	0.34	0.353
	(0.272)	(0.240)	(0.362)	
(leaders only)	15.72	14.39	1.34	0.020
	(0.285)	(0.468)	(0.557)	
Belief about relative performance (3)	58.24%	56.55%	-1.69	0.441
	(1.746)	(1.720)	(2.452)	
(leaders only)	70.04%	64.65%	-5.39	0.422
	(2.920)	(3.727)	(4.756)	
Belief about others performance (4)	12.62	12.31	0.31	0.306
	(0.205)	(0.226)	(0.306)	
(leaders only)	12.72	12.35	0.37	0.450
	(0.313)	(0.367)	(0.486)	
N	116	124	240	

Note: Table C.2 shows mean values for participants' performance, beliefs about absolute and relative performance, as well as beliefs about the average performance of others in the session. We report *p*-values of two-sample Mann-Whitney tests between treatments (i.e. Δ). Standard errors in parentheses.

C.2.2 Confidence at the Quartile Level

Table C.3: Average quartile overplacement by quartile and treatments.

	Q1 (top 25%)	Q2	Q3	Q4 (bot. 25%)
SI	-0.76	0.24	0.97	1.62
	(0.137)	(0.095)	(0.145)	(0.152)
AI	-0.81	0.16	0.74	1.71
	(0.142)	(0.115)	(0.146)	(0.133)

Note: Table C.3 reports the mean values for placement, measured as the belief about the percentage of participants outperformed and overplacement measured as the difference between placement and the actual proportion of participants outperformed by quartiles, for each treatment.

Table C.3 displays the average quartile overplacement measured as the difference between participant's belief about the quartile he thinks he belongs to (from 1, the bottom 25% to 4, the top 25%) and the quartile he actually belongs to (from 1, the bottom 25% to 4, the top 25%) in each treatment across quartiles. There-

fore, if the participant believes he is in the top 25% and the participant actually belongs to the top 25%, quartile overplacement will be equal to 0 (the participant is well-calibrated). On the contrary, if the participants believe he belongs to the top 25% but actually belongs in the bottom 25%, quartile overplacement will be equal to +3. Table C.3 shows that contrarily to what we expected, participants in the bottom half of the distribution of performance are more likely to be overconfident than participants in the top half of the distribution of performance (MW tests: $p < 0.001$ in both treatments).

C.2.3 Results on Beliefs

In order to further investigate the causal effect of the treatment on beliefs, we estimate OLS regressions in which we use beliefs about absolute performance as the dependent variable in models (1) and (2), beliefs about relative performance in models (3) and (4) and beliefs about the probability to be the top-ranked member of the group before the chat in models (5) and (6); and after the chat in models (7) and (8). The independent variable is a treatment dummy equals to 1 if the participant is in the AI treatment and 0 if the participant is in the SI treatment. We control for participants' score in models (1) and (2), and participants' actual percentile in model (3) to (8). In models (2), (4), (6) and (8) we also control for participants' demographics (sex, age, risk preferences and social dominance), as well as session size fixed effects.

Models (1) to (7) show no significant treatment effect on participants' beliefs. Model (8) shows a significant negative treatment effect on participants' beliefs about their probability to be the top-ranked member of their group after the chat. However, this effect is only observed after the chat. It is unclear whether it comes from the treatment as such or from its interaction with the communication of beliefs in the chat. Table C.8 provides support for this argument. Models (2), (4), (6) and (8) show that people who score high on the AMS scale (i.e. people who report themselves as more socially dominant) also hold higher beliefs about their absolute and relative performance, and the effect is significant at the 1% level. Finally, model (4) and (6) show that women tend to hold lower beliefs about their relative performance compared to men, and the effect is significant at the 1% and 10% levels respectively.[1]

[1]This results is consistent with Soldà et al. (2019) who find no difference in beliefs between gender regarding absolute performance but a significant effect on relative performance with women being less confident than men.

Table C.4: Treatment effect on participants' beliefs.

	belief about absolute perf.		belief about relative perf.		belief about being top-ranked (before the chat)		belief about being top-ranked (after the chat)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI	-0.170	-0.301	-1.584	-1.638	-0.219	-1.122	-3.739	-5.534**
	(0.330)	(0.320)	(2.315)	(1.934)	(3.157)	(2.937)	(2.952)	(2.563)
score	0.470***	0.473***						
	(0.075)	(0.069)						
percentile			0.225***	0.213***	0.217***	0.199***	0.222***	0.246***
			(0.040)	(0.042)	(0.055)	(0.057)	(0.065)	(0.067)
N=8		Ref.		Ref.		Ref.		Ref.
N=12		-0.347		1.235		0.422		-5.344
		(0.323)		(2.609)		(3.738)		(3.325)
N=16		0.411		3.727		6.711		0.167
		(0.401)		(2.989)		(4.059)		(3.424)
AMS score		0.099***		0.675***		1.372***		1.043***
		(0.036)		(0.248)		(0.371)		(0.387)
Female		-0.491		-8.490***		-5.312*		3.761
		(0.306)		(1.913)		(2.895)		(2.914)
Age		0.048		0.359*		0.102		0.512*
		(0.032)		(0.208)		(0.302)		(0.270)
Risk pref.		-0.047		-0.437		0.094		0.414
		(0.092)		(0.572)		(0.703)		(0.916)
Constant	8.327***	5.353***	47.983***	29.237***	47.138***	12.459	46.595***	6.107
	(1.008)	(1.562)	(2.588)	(9.181)	(3.337)	(12.844)	(4.076)	(12.512)
N	240	240	239	239	240	240	240	240

Note: Table C.4 shows the results of the OLS estimations of participants' beliefs on the treatment. Standard errors in parentheses. *** p <0.01, ** p <0.05, * p <0.10.

C.2.4 Results on Leaders' Selection

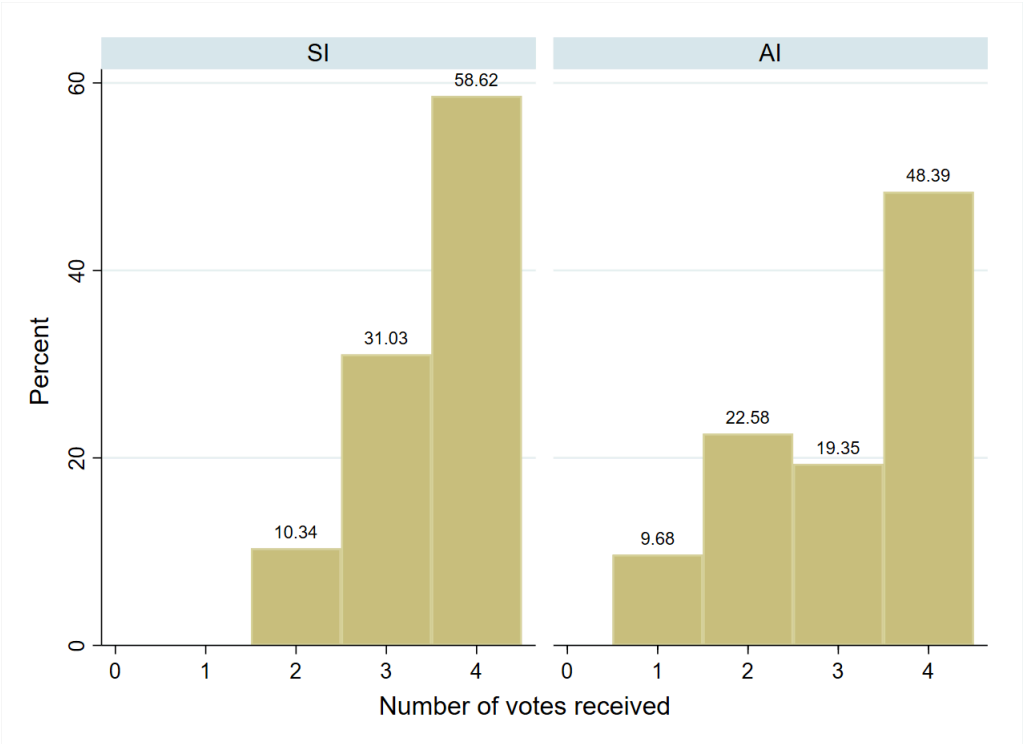


Figure C.2: Proportion of number of votes received by leaders, by treatments.

In our design, we allow participants to vote for themselves. One question that can arise is whether the difference in the ability to select the best performer of the group as a leader between treatments comes from a higher proportion of participants voting for themselves in the AI treatment. We find that participants vote more for themselves in the AI treatment than in the SI treatment but the difference is only marginally significant (33.62% vs. 44.35% two-sample test of proportion: p =0.089). However, 3 leaders were determined randomly (i.e. all group members voted from themselves) in the AI treatment while this case never happened in the SI treatment. Figure C.2 displays the distribution of the number of votes received by leaders in both treatments. In both treatments, leaders received the majority of votes from their group members (89.65% in the SI treatment and 67.74% in the AI treatment). This result suggests that most leaders successfully convinced their fellow group members that they were the top-ranked performer.

Table C.5: Determinants of votes.

Dep. variable:	pooled		SI		AI	
Nb. of votes received	(1)	(2)	(3)	(4)	(5)	(6)
Belief percentile	0.025***	0.024***	0.031**	0.033**	0.021**	0.016*
	(0.007)	(0.008)	(0.013)	(0.014)	(0.009)	(0.010)
percentile	0.010**	0.010**	0.017**	0.017**	0.004	0.004
	(0.005)	(0.005)	(0.007)	(0.007)	(0.006)	(0.006)
N=8		Ref.		Ref.		Ref.
N=12		-0.040		-0.178		0.068
		(0.108)		(0.154)		(0.167)
N=16		-0.100		-0.286		0.014
		(0.112)		(0.204)		(0.154)
AMS score		0.044**		0.028		0.064***
		(0.019)		(0.027)		(0.024)
Female		0.026		0.229		-0.133
		(0.183)		(0.257)		(0.287)
Age		-0.008		-0.025		-0.001
		(0.016)		(0.027)		(0.019)
Risk pref.		-0.035		-0.013		-0.056
		(0.044)		(0.061)		(0.058)
Constant	-2.083***	-2.664***	-2.923***	-2.998***	-1.467***	-2.391**
	(0.395)	(0.759)	(0.668)	(1.101)	(0.463)	(1.118)
Obs.	239	239	116	116	123	123

Note: Table C.5 shows the results of the Poisson estimation of the number of votes received by participant i on participant i's belief about their relative performance. Standard errors in parentheses. *** p <0.01, ** p <0.05, * p <0.10.

Table C.5 reports Poisson regressions in which the dependent variable is the number of vote received by participant i. The independent variables include participant's belief about the percentage of participants he outperformed. Models (1) and (2) consider all the observations. Models (3) and (4) consider the observations for the SI treatment only. Models (5) and (6) consider the observations for the AI treatment only. We control for the percentage of participants outperformed in models (1), (3) and (5), and participants demographics (sex, age, risk preferences and social dominance) as well as session size fixed effects in models (2), (4) and (6).

Models (1) to (5) shows a positive effect of beliefs on the number of votes received by participants and these results are significant at the 5% results. The effect is still positive in model (6) when controlling for participants' demographics in the AI treatment but the effect is only marginally significant. Interestingly, participants' social dominance seems to positively affect the number of votes received but this effect is only significant in the AI treatment.

C.2.5 Results on Leaders' Selection (Quartile Level)

In columns (1) and (2), we estimate logistic regressions in which the dependent variable is a leader dummy that equals 1 if the participant was chosen as the leader and 0 otherwise. In columns (3) and (4), we estimate of Poisson regressions in which the dependent variable is the number of votes received by participant i. Independent variables for models (1) to (4) include a treatment dummy and quartile dummies (top 25% (Q1); top 50% (Q2); bottom 25% (Q4) and bottom 50% (Q3)), as well as the interaction terms between the two. The marginal effects are displayed in Table C.6. In model (2) and (4), we control for session size fixed-effect, as well as participants' demographics (sex, age, risk preferences and social dominance).

Models (1) and (2) show that participants from the top 25% are more likely to be chosen as leaders in the SI treatment. Models (3) and (4) shows that these participants are also more likely to receive the more votes and these results are significant at the 5% level. Interestingly, models (1) and (2) show that in the AI treatment, participants from the top 50% are more likely to be chosen as leaders than participants from the top 25% and this effect is significant at the 5% level. Model (3) and (4) display a similar pattern but the results are only marginally significant.

Table C.6: Treatment effect on leaders' selection between quartiles.

Dep. variable:	Leader		Nb. of votes received	
	(1)	(2)	(3)	(4)
AI	-0.193**	-0.191**	-0.403	-0.391
	(0.094)	(0.094)	(0.253)	(0.250)
Q1	Ref.	Ref.	Ref.	Ref.
Q2	-0.283***	-0.283***	-1.310***	-1.306***
	(0.069)	(0.069)	(0.447)	(0.430)
Q3	-0.258***	-0.257***	-1.008**	-1.009**
	(0.073)	(0.072)	(0.425)	(0.425)
Q4	-0.313***	-0.299***	-1.578***	-1.450***
	(0.066)	(0.068)	(0.527)	(0.524)
AI*Q1	Ref.	Ref.	Ref.	Ref.
AI*Q2	0.481**	0.472**	1.000*	0.953*
	(0.220)	(0.222)	(0.590)	(0.576)
AI*Q3	0.262	0.292	0.246	0.339
	(0.245)	(0.247)	(0.581)	(0.578)
AI*Q4	0.350	0.321	0.990	0.896
	(0.259)	(0.263)	(0.671)	(0.652)
N=8		Ref.		Ref.
N=12		0.009		0.058
		(0.017)		(0.058)
N=16		0.013		0.069
		(0.014)		(0.052)
AMS score		0.013**		0.060***
		(0.006)		(0.019)
Female		-0.051		-0.139
		(0.051)		(0.168)
Age		0.001		-0.004
		(0.005)		(0.016)
Risk pref.		-0.005		-0.029
		(0.015)		(0.041)
Constant	–	–	0.776***	-0.446
			(0,152)	(0.669)
N	240	240	240	240

Note: Columns (1) and (2) show the marginal effects of logit regressions of the likelihood to be selected as the leader on the interaction between treatments and quartiles. Columns (3) and (4) show the estimates of the Poisson regressions of the number of votes received on the interaction between treatments and quartiles. Standard errors in parentheses. *** p <0.01, ** p <0.05, * p <0.10.

C.3 Communication

C.3.1 Chat Content Analysis

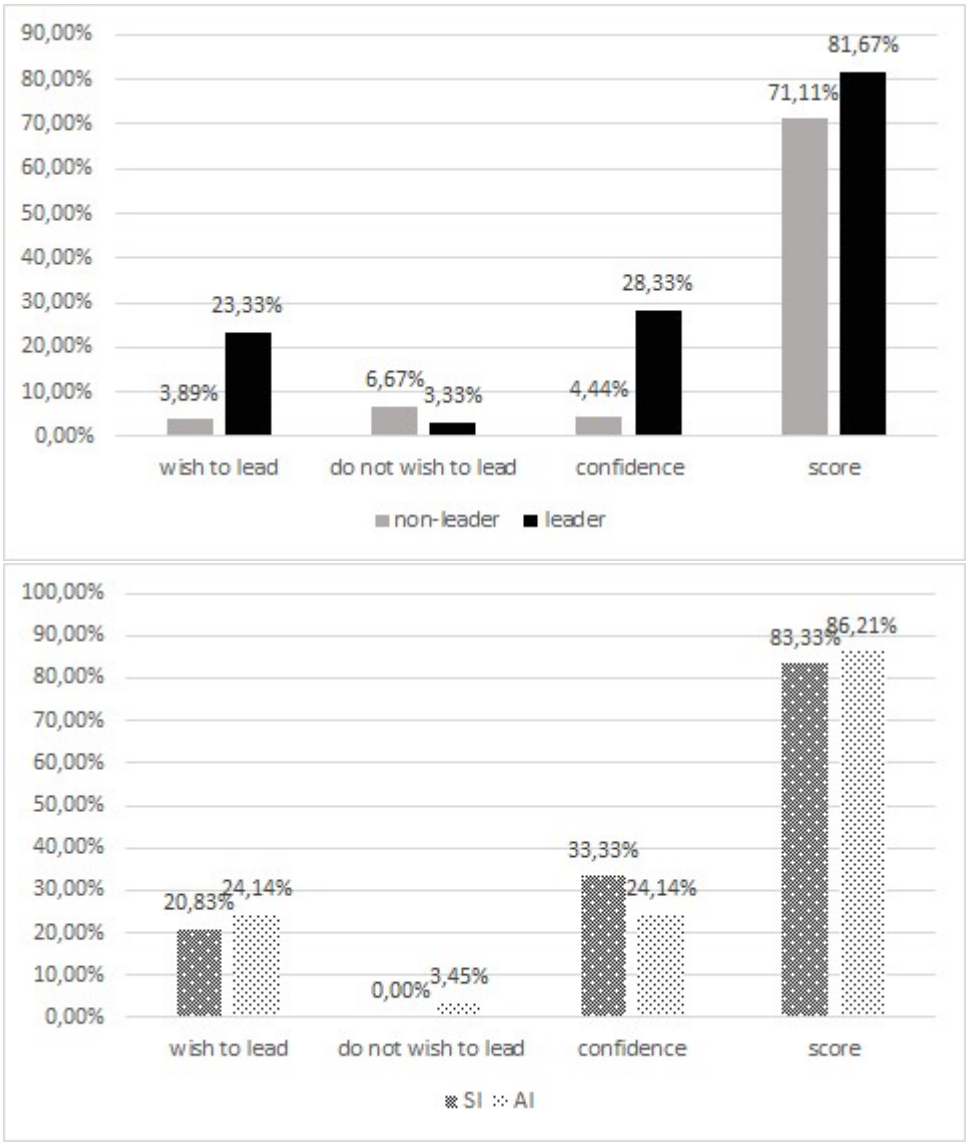


Figure C.3: Summary of chat messages for both leaders and non-leader (upper panel) and leaders only, by treatments (lower panel).

In our experiment, participants are allowed to communicate with their group members via an interactive chat box for at most ten minutes. At the end of these ten minutes, one group member will be chosen as the leader of the group. In this section, we first provide an analysis of the content of the chat. For each group member, we record (i) whether the participant gives an estimate of his performance at the Raven's matrices (if this is the case, we also record the number); (ii) a dummy variable equals to 1 if the group member expresses his or her willingness to be the leader of the group and 0 otherwise; (iii) a dummy variable equals to 1

if the group members expresses his or her willingness NOT to be the leader of the group; and (iv) a dummy variable equals to 1 if the group member is defined by at least one other group member as confident. For each group, we also created a dummy variable equals to 1 if the group agree on using a randomization device to select the leader of the group; and 0 otherwise.

The upper panel of Figure C.3 summarizes the chat messages belonging to each of the category mentioned above for leaders and followers, and the lower panel displays summarizes the chat messages for leaders only, by treatments. Since the lower panel of Figure C.3 shows no significant differences in the messages content between treatments, we will focus our analysis between leaders and non-leaders. The upper panel of Figure C.3 shows that leaders express their willingness to lead more often than followers and are also more likely than followers to be described as confident by at least one of their group member (two-sample tests of proportion: $p < 0.001$ in both cases). Interestingly, being identified as confident is positively correlated with participants' beliefs about their absolute performance (Spearman correlation: $r_s = 0.229$, $p = 0.002$) but not their actual score (Spearman correlation: $r_s = 0.036$, $p = 0.639$).

To investigate the effect of the chat messages content on the likelihood to become a leader, we estimate Poisson regressions in which the dependent variable is the number of vote received by participant i. The independent variables include participant i's belief about his or her probability to be the top-ranked member of the group, participant i's performance at the Raven's matrices, a dummy variable equals to 1 if participant i was described as confident by at least one member of the group, and 0 otherwise; a dummy variable equals to 1 if participant i clearly expressed is willingness to be the leader of the group, and 0 otherwise and participant i's social dominance.

Results are reported in Table C.7. Models (1) and (2) consider all the observations. 177 out of 240 participants reported an estimate of their score at the Raven's matrices in one of their chat messages. In models (3) and (4), we only consider participants who mention an estimate of their score during the chat and add this reported estimate as an independent variable. In models (2) and (4), we control for participant i demographics (sex, age, risk preferences, social dominance and whether English is is participant i native language).

Table C.7: Determinants of votes.

Dep. variable:	Nb. of votes received			
	(1)	(2)	(3)	(4)
Reported score	–	–	0.223***	0.222***
			(0.052)	(0.055)
Belief top-ranked	0.013***	0.013***	0.006*	0.007*
	(0.003)	(0.003)	(0.003)	(0.004)
performance	0.111***	0.096***	0.061*	0.050
	(0.027)	(0.028)	(0.032)	(0.033)
Confident	0.525***	0.522***	0.507***	0.524***
	(0.160)	(0.163)	(0.172)	(0.174)
Wish to lead	0.732***	0.791***	0.688***	0.748***
	(0.160)	(0.165)	(0.175)	(0.184)
AMS score	0.027	0.029*	0.035*	0.037*
	(0.017)	(0.017)	(0.021)	(0.021)
English		0.382***		0.260*
		(0.133)		(0.151)
Female		-0.053		-0.021
		(0.133)		(0.150)
Age		-0.008		-0.015
		(0.013)		(0.015)
Risk preferences		-0.029		-0.012
		(0.032)		(0.037)
Constant	-3.028***	-2.734***	-5.585***	-5.215***
	(0.524)	(0.679)	(0.931)	(1.000)
Obs.	240	240	177	177

Note: Table 4.4 shows the results of the Poisson estimation of the number of votes received by participant i on participant i's chat messages content. Standard errors in parentheses. *** p <0.01, ** p <0.05, * p <0.10.

Model (1) to (4) in Table C.7 shows a strong positive effect of being described as confident by at least one other group member and expressing one's willingness to be the leader of the group on the number of votes received. In model (3) to (4), we also find a strong positive relationship between the estimate of one's score reported in the chat and the number of votes received. These results are in line with the idea of strategic overconfidence: participants who hold high beliefs about their performance seem able to convey their confidence to their group members more effectively and are selected as the leader of their group more often.

C.3.2 Effect of Communication on Beliefs

In our experiment, participants do not receive any feedback about their performance. However, participants can also use the chat to infer information about their group members. This information, can in turn, that can affect their beliefs

about their relative performance. We now analyze the effect of communication on participants beliefs about their likelihood to be the best performer in the group after the chat. Figure C.4 displays participants beliefs before the chat (light bars) and after the chat (dark bars) for participants in the top 25% and participants not in the top 25%, in both treatments.

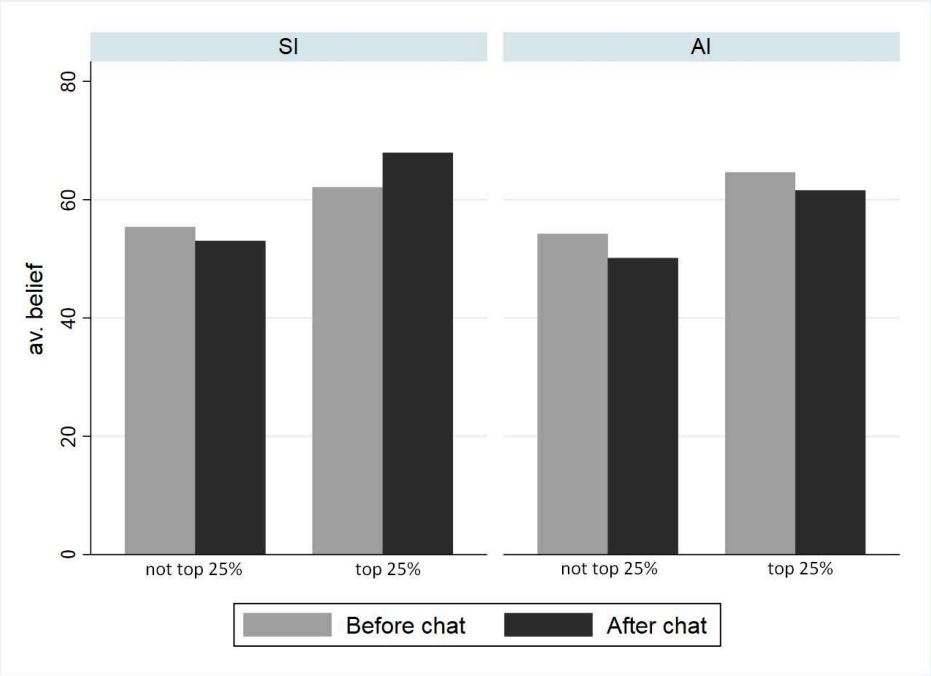


Figure C.4: Average belief before and after the chat for participants in the top 25% and participants not in the top 25%, in both treatments.

Figure C.4 shows that participants in the SI treatment update their beliefs in the correct direction: participants in the top 25% update upwards while participants not in the top 25% update downwards. In the AI treatment, all participants update their beliefs downwards after the chat. These results suggest that the chat provides some useful information in the SI treatment but not in the AI treatment.

To further investigate the effect of the treatment on belief updating, we estimate OLS regressions in which the dependent variable is participants' beliefs after the chat. Independent variables include a categorical variable of participants' quartiles, a treatment dummy and an interaction term. The results are displayed Table C.8. In model (1), we control for participants' beliefs before the chat. In model (2), we also control for participants' demographics (sex, age, risk preferences, social dominance and whether English is their native language).

Table C.8: Treatment effect on belief after the chat.

Dep. variable:	Beliefs after the chat	
	(1)	(2)
Treatment	-7.951	-8.657
	(5.326)	(5.280)
Prior	0.638***	0.640***
	(0.055)	(0.056)
Q1	ref.	ref.
Q2	-9.805*	-10.817**
	(5.419)	(5.420)
Q3	-14.742***	-15.978***
	(5.415)	(5.382)
Q4	-7.280	-10.860**
	(5.450)	(5.516)
Treat. x Q1	Ref.	ref.
Treat. x Q2	3.638	2.653
	(7.530)	(7.465)
Treat. x Q3	11.015	11.149
	(7.537)	(7.446)
Treat. x Q4	2.709	4.335
	(7.540)	(7.495)
AMS score		0.296
		(0.330)
Female		7.391***
		(2.720)
Age		0.397*
		(0.233)
Risk preferences		0.619
		(0.680)
English		0.475
		(2.775)
Constant	28.330***	4.972
	(5.137)	(11.050)
Obs.	240	240

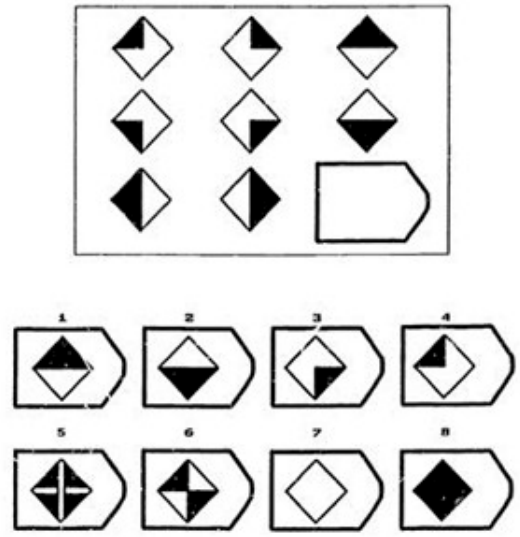
Note: Table C.8 displays the OLS estimations of posterior beliefs about relative performance on the interactions between the treatment and participants' actual quartile. Standard errors in parentheses. * $p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

Results from Table C.8 support our previous findings and show that in the SI treatment, participants below the top 25% update their beliefs downwards compared to participants in the top 25% (Note that the effect is mainly driven by participants in the bottom 50% but not in the bottom 25%). We do not find such effect in the AI treatment.

C.4 Instructions

INSTRUCTIONS PART 1

In this part, you will be asked to solve 20 Raven's matrices. Raven's matrices are multiple choice intelligence tests. They measure the capacity to reason clearly and understand complexity. Here is an example of a Raven's matrix:



For each matrix, you have to select the correct answer among 8 possibilities. In the example above, the correct answer is 8.

WARNING: Once you select an answer, the next matrix will be automatically displayed on your screen and you won't be able to go back to change your answer.

For each matrix, you will have 60 seconds. At the top of your screen, you will have a counter which continuously displays the remaining time. The last 10 seconds of the remaining time will be displayed in red to further urge you to make a decision quickly.

It is important that you make your choice within the allocated time. If you do not make a choice after 60 seconds, the computer will automatically skip to the next matrix.

Payoffs
You will receive 0.5 AUD for each matrix solved correctly. You will not lose nor earn anything for the incorrect matrix. Your payoff from this part and the total number of matrices you have correctly solved will be revealed to you at the end of

today's experiment. Before beginning the first Part, you will be able to practice on two matrices which will not be paid. You can use as much time as you want to answer these two matrices.

* * *

If you have any question, please raise your hand and the experimenter will answer you in private. Once all the questions have been answered, the practice rounds will start.

INSTRUCTIONS PART 2 (16 participants version)

Group formation

In this part, you will be assigned to a group of 4 participants. The groups will be formed according to the following procedure: All participants in this room will be ranked according to the score they received in Part 1. The participant with the highest score in Part 1 is ranked 1 and the participant with the lowest score in Part 1 is ranked 16 (Note that there are 16 participants in the room). In case of ties between several participants, the computer will randomly allocate the corresponding ranks to these participants. For example, if participants 7 and 11 both have solved the highest number of matrices in part 1, the computer will randomly put one in rank 1 and the other one in rank 2. After all the participants are ranked, the rankings will be divided into 4 categories as shown in the table below:

Category	Definition	Participant Rank
Q1	Participants ranked as the top 25% of all participants in the session.	1 2 3 4
Q2	Participants ranked below the top 25% group (Q1), but in the top 50% of all participants in the session.	5 6 7 8
Q3	Participants ranked above the bottom 25% group (Q4), but in the bottom 50% of all participants in the session.	9 10 11 12
Q4	Participants ranked in the bottom 25% of all participants in the session	13 14 15 16

The computer will then randomly draw one participant from each category to form a group, until all four groups are formed. This means: for each group

formed by the computer there should be one participant from Q1, one participant from Q2, one participant from Q3, and one participant from Q4. You will not be told your rank or the rank of the other participants.

Leader selection
Once groups are formed, your task is to choose a leader within your own group. The role of the leader is to make decisions that will affect the payoffs of everyone in the group. The group will benefit more if the leader your group selected belongs to category Q1 (i.e., is the best performer in your group). The leader will also receive a bonus for being the leader.

You will be allowed to communicate with your fellow group members for 10 minutes as you make this decision via a group chat box. After the chat, you will nominate a leader anonymously. The group member who receives the highest number of votes is assigned as the leader for the group. In case of ties; the computer program will randomly select one of the participants with the most votes to be the leader. After all group members have cast a vote, the role of each group member (either as leader or as follower) will be displayed on the screen. You will not be told how many votes each group member received. Following the election outcome, the leader will then make a series of decisions, one of which will be randomly selected for payment. You will be given more details regarding these decisions after your group has selected a leader.

The series of decisions that the leader will make in Part 2 is different from the decisions in Part 1 and does not involve solving Raven's matrices.

Payoffs
Your payoffs for the second part of the experiment are composed of two components:

- One component is based on the selected leader's performance in Part 1: you will receive 0.5 AUD per Raven's matrices solved correctly by your leader in Part 1.
- The other part is based on the leader's decision: everyone in your group will receive between 0 and 10 AUD depending on the leader's decisions. Your chances to receive a higher payment will depend partially on whether the leader your group selected was in in the highest performing category Q1 (i.e., was the best performer in your group).

(AI treatment only) On top of that, the leader will receive a bonus of 5 AUD

for being the leader.

To sum up, your total payoff for part 2 is the following:

- If you are a leader, you will receive:
  (AI treatment) 5 AUD (leader bonus) + 0.5 AUD times the number of matrices solved correctly by you in part 1 + the number of AUD generated by your decision in Part 2.
  (SI treatment) 0.5 AUD times the number of matrices solved correctly by you in part 1 + the number of AUD generated by your decision in Part 2.
- If you are a follower, you will receive: 0.5 AUD times the number of matrices solved correctly by your leader in part 1 + the number of AUD generated by your leader's decision in Part 2.

In summary:

- The computer program will form groups of 4 participants according to people's performance in Part 1.
- You will be given the opportunity to anonymously chat with your group member for 10 minutes during which you will select a leader.
- The leader will make a series of decision and the outcome of these decisions will be affected by the likelihood that the leader is the best performer in your group.
- Your payoffs in part 2 will depend both on the leader's performance in Part 1 and the leader's decisions in part 2.
- **(AI treatment only)** The person chosen as leader will receive a bonus of 5 AUD on top of other payoffs.

* * *

If you have any question, please raise your hand and the experimenter will answer you in private. Once we answer all your questions, a comprehension questionnaire will be displayed on your screen.

INSTRUCTIONS PART 2 - Leader's Decisions

For this part of the experiment, the leader needs to make 10 choices. The decisions are displayed in the table below:

Decision	Option A	Option B
1	\$10 if you are the best performer in your group; \$0 if not	\$1
2	\$10 if you are the best performer in your group; \$0 if not	\$2
3	\$10 if you are the best performer in your group; \$0 if not	\$3
4	\$10 if you are the best performer in your group; \$0 if not	\$4
5	\$10 if you are the best performer in your group; \$0 if not	\$5
6	\$10 if you are the best performer in your group; \$0 if not	\$6
7	\$10 if you are the best performer in your group; \$0 if not	\$7
8	\$10 if you are the best performer in your group; \$0 if not	\$8
9	\$10 if you are the best performer in your group; \$0 if not	\$9
10	\$10 if you are the best performer in your group; \$0 if not	\$10

For each decision, there are 2 options:

- (Option A): Receiving \$10 if the leader's performance in Part 1 was the best in your group and \$0 if it was not.
- (Option B): Receiving a fixed amount of money independently of the leader's performance in Part 1.

While Option A always stays the same, the fixed earning in Option B increases incrementally.

The leader's task is to pick *either Option A or Option B* for every decision.

For example, choosing Option B in decision 5 means that the leader prefers everyone in the group to receive 5 AUD regardless of his or her performance in Part 1, instead of receiving 10 AUD if he or she was the best performer in the group in Part 1.

To maximize the chance to receive the highest amount as possible, the leader should switch from Option A to Option B depending on his/her beliefs about his/her probability to be the best performer in the group.

The farther down the table the leader switches from Option A to Option B, the more confident he or she is that he or she is the best performer in the group.

At the end of the task, one of the leader's decisions will be randomly selected for payment. If the leader chose Option B in the selected decision, everyone in the group will receive the amount of AUD proposed in Option B. If the leader chose

Option A in the selected decision, everyone in the group will receive 10 AUD if the leader's performance in Part 1 was the best in the group, otherwise everyone in the group will receive 0 AUD from this task.

C.5 Screenshots Belief Elicitation

Figure C.5 to C.8 are screenshots of the instructions of the belief elicitation. Figure C.5 shows the instruction for our belief elicitation about absolute performance. Figure C.6 shows the instruction for our belief elicitation about absolute performance of other participants in the session. Figure C.7 shows the instruction for our belief elicitation about relative performance. Finally, Figure C.8 shows the instruction for our belief elicitation about the probability that the participant belong to the top 25%, both before and after the chat.Note that the instructions in Figures C.6 and C.7 are different depending on the number of participants n in the session. Here, we provide the instructions for a session with 16 participants.

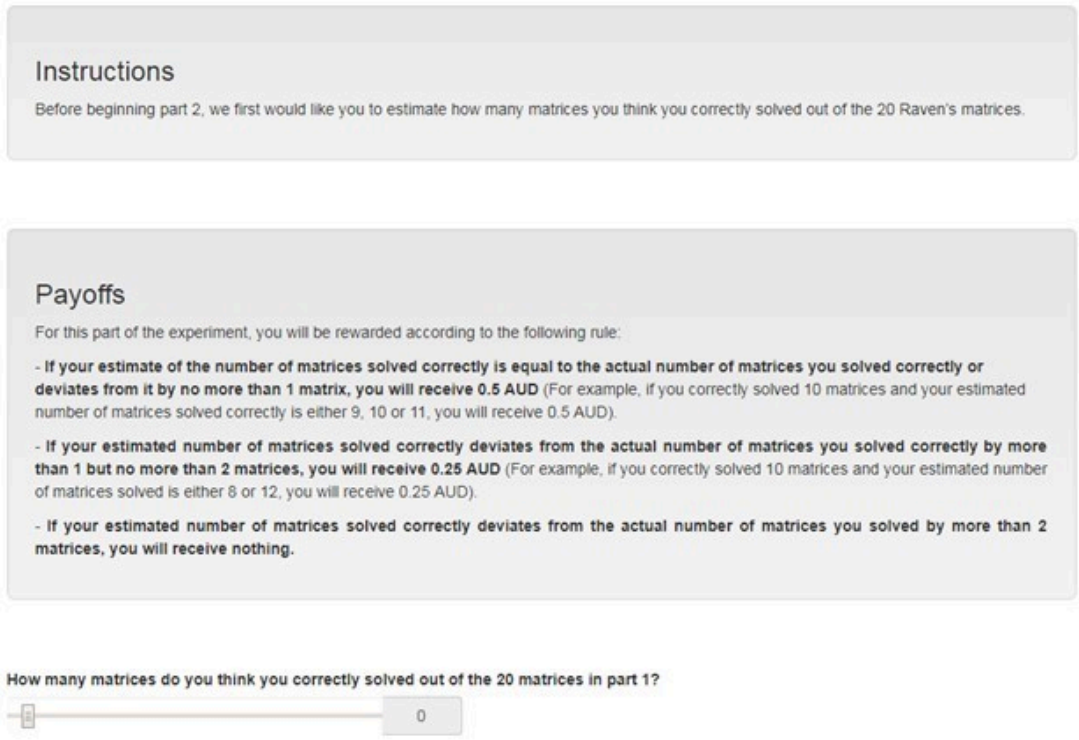


Figure C.5: Belief elicitation about performance

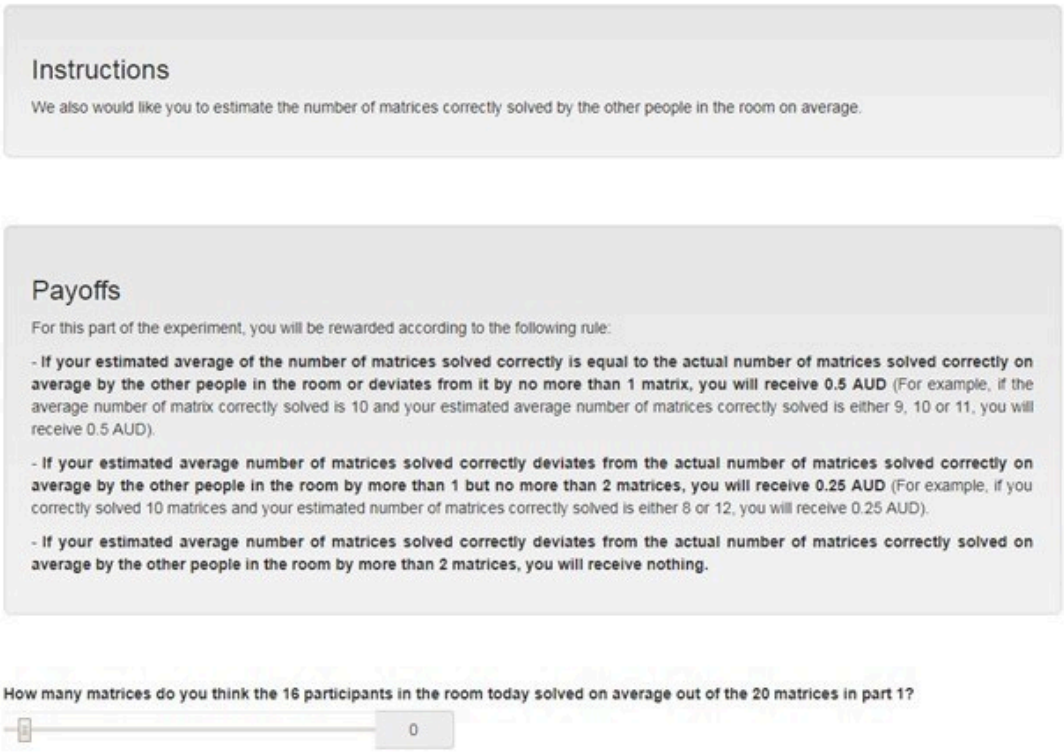


Figure C.6: Belief elicitation about others' performance

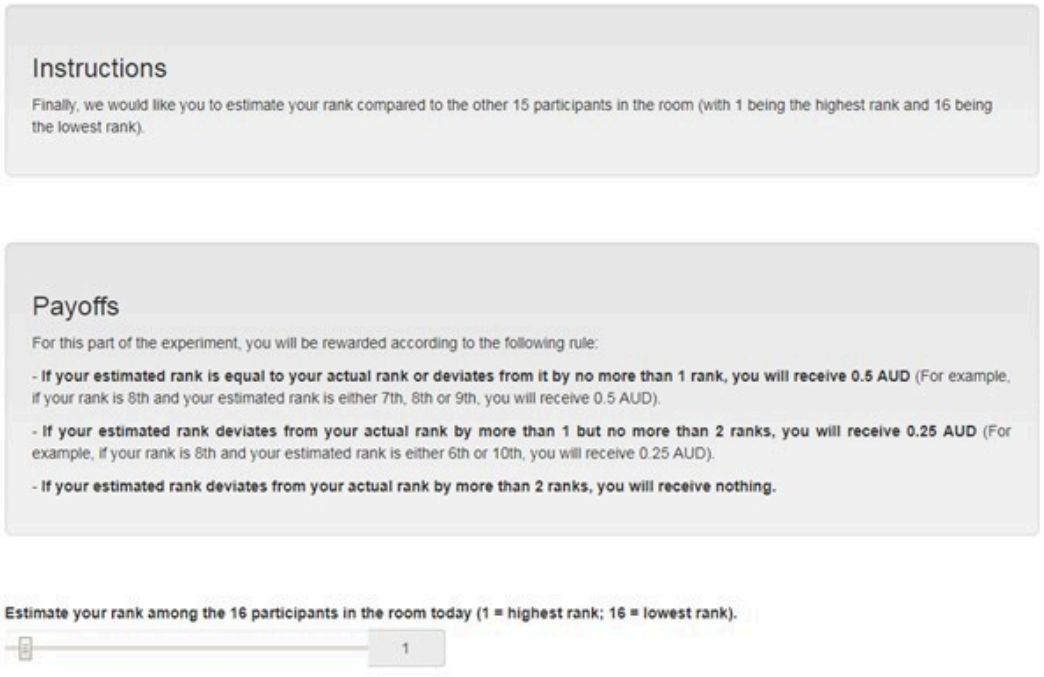


Figure C.7: Belief elicitation about relative performance

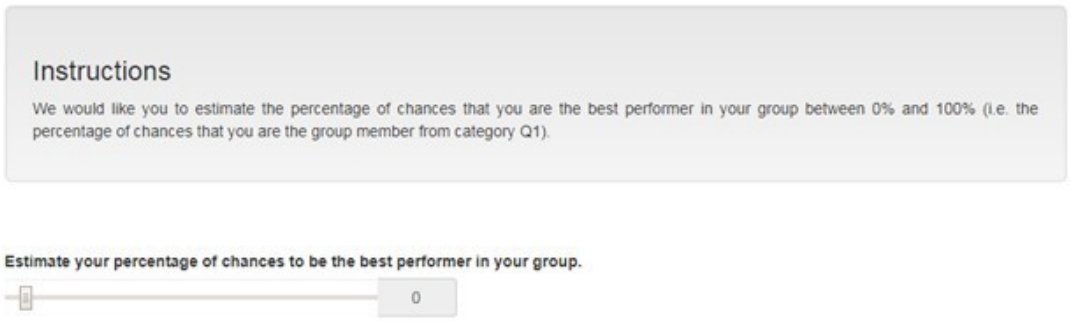


Figure C.8: Belief elicitation about the probability to be in the top 25% (before and after the chat)