CESifo WORKING PAPERS

7517
2019

February 2019

On Self-Serving Strategic Beliefs

Nadja R. Ging-Jehli, Florian H. Schneider, Roberto A. Weber

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 7517*
Category 13: Behavioural Economics

On Self-Serving Strategic Beliefs

Abstract

We experimentally study settings where an individual may have an incentive to adopt negative beliefs about another's intentions in order to justify egoistic behavior. Our first study uses a game in which a player can take money from an opponent in order to prevent the opponent from subsequently causing harm. We hypothesize that players will justify taking by engaging in "strategic cynicism," convincing themselves of the opponent's ill intentions. We elicit incentivized beliefs both from players with such an incentive and from neutral third parties with no incentive to bias their beliefs. We find no difference between the two sets of beliefs, suggesting that people do not negatively bias their beliefs about a strategic opponent even when they have an incentive to do so. This result contrasts with Di Tella, et al. (2015), who argue that they provide evidence of strategic cynicism. We reconcile the discrepancy by using Di Tella, et al.'s, data, a simple model of strategic belief manipulation and a novel experiment in which we replicate Di Tella, et al.'s, experiment and also elicit the beliefs of neutral third parties. Across three experimental datasets, the results provide no evidence of negatively biased beliefs about others' intentions. However, Di Tella, et al.'s, results and our novel data indicate that those with a greater incentive to view others' intentions negatively exhibit relatively less positive beliefs than those without such incentives.

JEL-Codes: C720, D830, C920.

Keywords: motivated beliefs, strategic cynicism, bias, experiment.

Nadja R. Ging-Jehli
The Ohio State University
Columbus / Ohio / USA
gingjehli.1@buckeyemail.osu.edu

Florian H. Schneider
University of Zurich
Zurich / Switzerland
florian.schneider2@econ.uzh.ch

*Roberto A. Weber**
Department of Economics
University of Zurich
Blümlisalpstrasse 10
Switzerland – 8006 Zurich
roberto.weber@econ.uzh.ch

*corresponding author

January 23, 2019
We thank participants at the 2015 Self-Deception, Self-Signaling and Self-Control Workshop in Toulouse, the 2016 CUSO Workshop in Freiburg, the 2016 ZurichWorkshop in Economics, the Barcelona GSE seminar and the 2018 CESifo Behavioural Economics Conference. We thank Roland Bénabou, Botond Kőszegi, Ricardo Perez-Truglia, Martin Schonger, Ivo Schurtenberger and Christian Zehnder for valuable comments.

1. Introduction

Considerable evidence indicates that decision makers confronted with tradeoffs between egoistic and social considerations, such as fairness and equality, will rely on justifications to prioritize the former while avoiding the impression that they are acting selfishly (Dana, Weber and Kuang, 2007; Hamman, Loewenstein and Weber, 2010; Gino, Norton and Weber, 2016; Grossman and van der Weele, 2017). This includes engaging in self-serving belief manipulation, whereby actions that are personally beneficial can be justified by changing one's beliefs or perceptions of what is fair (Babcock and Loewenstein, 1997; Konow, 2000), a product's quality (Chen & Gesche, 2017; Gneezy, Saccardo, Serra-Garcia, and van Veldhuizen, 2018) or the likely outcomes of a random process (Haisley and Weber, 2010; Exley, 2016).[1]

One important, but largely unexplored, context for self-serving belief manipulation is in strategic settings where individuals form beliefs about an opponent's likely behavior. It has long been recognized that beliefs about other players' actions and intentions can play a central role in prosocial behavior, with a positive relationship between the belief that others will act unkindly and one's own egoistic behavior (e.g., Rabin, 1993; Levine, 1998; Fischbacher, Gächter and Fehr, 2001). Strategic beliefs are typically assumed to be determined by the structure of the game and beliefs about others' preferences or rationality. However, in light of the apparent ease with which people bias their beliefs in self-serving ways in other contexts, it seems plausible that they may similarly bias their beliefs about others' actions when doing so can justify acting in a selfish way that harms others.[2] Indeed, a recent paper by Di Tella, Perez-Truglia, Babino and Sigman (2015) provides evidence consistent with the idea that people engage in such "strategic cynicism." Specifically, they

[1] Such self-serving interpretations are related to the concept of "motivated reasoning" from psychology (Kunda, 1990). Models introducing self-deception and self-image concerns to economics include Akerlof and Dickens (1982), Rabin (1994), Akerlof and Kranton (2000), Bénabou and Tirole (2006, 2011), Bénabou (2013). There is also evidence for self-serving belief manipulation about other desired qualities, like one's abilities (Möbius, Niederle, Niehaus and Rosenblat, 2017; Zimmerman, 2018), beauty (Eil and Rao, 2011), honesty (Mazar, Amir, and Ariely 2008; Shalvi, Gino, Barkan, and Ayal, 2015) and about desired future life events (Irwin, 1953; Mayraz, 2013).

[2] There are many contexts in which adopting cynical beliefs about others' likely actions may be strategically desirable for an individual constrained to act morally. For instance, an employer who can benefit by laying off a worker may find it easier to do so if she adopts the belief that the employee is likely committing acts that merit firing. A national leader intent on seizing land from a neighboring country may find this easier to justify under the belief that the other country intends to act aggressively. A US President may find it easier to justify firing a special counsel investigating him for misconduct if he convinces himself that the investigation is a WITCH HUNT! Trivers (2011) discusses an alternative motive to engage in self-deception in strategic situations: deceiving oneself about one's own qualities, such as ability, might be an effective strategy to deceive others (see also Schwardmann and van der Weele (2016), for related experimental evidence).

demonstrate that people with a greater opportunity to take from another person believe that this opponent is more likely to act in a greedy and harmful manner.

Our study investigates the phenomenon of strategic self-deception, although we initially approach this question in a different manner from Di Tella, et al. Rather than testing whether people with a *greater* incentive to take from others adopt *relatively* more negative beliefs about these opponents, as they do, our focus is on whether people with the opportunity to take from others adopt beliefs that are biased in comparison to two reasonable objective standards: the actual empirical behavioral frequency of opponents' behavior and the beliefs of neutral outsiders with no incentive to view others self-servingly. That is, we test the extent to which individuals with an incentive to engage in strategic cynicism adopt beliefs that are negatively biased in *absolute* terms. In contrast, Di Tella, et al., study a *relative* form of this bias, investigating whether one group's beliefs are more negative—or, critically, less positive—than those of another group.

Our results show that this distinction is important. We find evidence consistent with the relative bias documented by Di Tella, et al. However, we also show in two novel studies and in Di Tella, et al.'s, own data that there is no evidence of negatively biased strategic beliefs in absolute terms. In fact, across all three studies, individuals with an incentive to take from others—and, therefore, with an incentive to engage in strategic cynicism—actually hold highly accurate beliefs that are close to both the actual behavior of their counterparts and to the beliefs of neutral observers. The only bias, relative to these objective and neutral standards, in all three data sets lies in the beliefs of individuals in Di Tella, et al.'s, design with a *low* ability to take from their counterpart; these people exhibit overly *positive* beliefs about their counterpart's likely action. Thus, to the extent that absolute bias exists in people's beliefs about a counterpart's actions, it appears to be one of positivity rather than cynicism.

Our first study, which we conducted prior to knowing about Di Tella, et al.'s, related work, uses a game that we refer to as the "pre-emptive taking game." In this game, a pair of players—say, "Ann" and "Bob"—both start off with the same wealth endowment. Ann first decides how much to take from Bob. Bob then decides how much to take from Ann. A key feature is that Bob's ability to take from Ann increases in the amount of money he has remaining after Ann's taking decision. That is, by taking from Bob, Ann both increases her earnings and reduces his opportunity to act in a selfish and harmful manner. Thus, Ann's taking decision may naturally be influenced by whether or not she thinks Bob will use his remaining money to harm Ann. But, if Ann feels constrained to act "fairly," the game also

creates an incentive for Ann to manipulate her beliefs about Bob's likely action, since this gives her a justification for taking more under the guise of self-protection.

Our main purpose in this study is to test for a bias in the beliefs of subjects in the role of Ann. Therefore, we directly elicit such beliefs about the amount that Bob will take if given the opportunity. We compare this to beliefs elicited from neutral third parties who have no incentive to engage in strategic cynicism. Our hypothesis is that Ann's incentive to justify taking by adopting a cynical belief about Bob's likely behavior will lead her to self-servingly negatively bias these beliefs. Surprisingly, however, in light of other instances in which people seem to engage in self-deception in non-strategic contexts, we find no difference between the two sets of beliefs. The two beliefs are virtually identical and very close to the true empirical frequencies. This suggests, at the least, that there are limitations in people's ability to manipulate their beliefs about a strategic opponent.

This finding also contrasts with those of Di Tella, et al., who argue that their experimental evidence shows that individuals form biased beliefs and convince themselves that their counterparts are more likely act egoistically than they actually are.[3] Di Tella, et al., support this conclusion with an experiment using a game, labeled the "corruption game," that shares features with our pre-emptive taking game. In this game, Ann and Bob again start with an identical number of tokens and Ann similarly decides how many tokens to take from Bob. In the corruption game, however, Bob simultaneously makes a binary decision whether to act "corruptly," by taking a side payment that increases Bob's wealth while lowering the value of the tokens. Thus, Ann can justify taking more tokens from Bob as a fair action if she thinks that he will act corruptly. The experiment manipulates Ann's ability to take tokens from Bob and finds that subjects in the role of Ann adopt more pessimistic beliefs when they have the ability to take more tokens. Thus, individuals seem to respond to the incentive to take more from their counterpart by engaging in strategic cynicism.

At first, these two results seem to offer conflicting evidence. Our first study finds no evidence of strategic cynicism. In contrast, Di Tella, et al., conclude that people engage in self-serving belief manipulation. We reconcile this apparent inconsistency using the data from Di Tella, et al.'s, experiment, a simple model of strategic self-deception and an additional novel experiment. We show that the two sets of results are actually highly

[3] Specifically, in a context where the actual proportion of egoistic counterparts is p_0, Di Tella, et al., argue that a self-servingly biased decision maker "may form a biased belief [...] instead of correctly remembering a proportion of p_0 of low-type, the individual may try to convince herself that the proportion was actually $p > p_0$" (p. 3437).

consistent and, in doing so, provide important evidence on the nature of strategic self-deception.

First, we show that a closer inspection of Di Tella, et al.'s, data reveals that their results, like ours, do not actually show any absolute cynicism or bias on the part of individuals with an incentive to act egoistically. In fact, the subjects in their experiment with the greater incentive to engage in strategic self-deception provide belief estimates very close to the actual frequency of egoistic behavior by their counterparts. In contrast, those subjects with a low incentive to engage in strategic cynicism exhibit the most bias, but in the direction of believing that their opponents will be *less egoistic* than they actually are. Thus, to the extent that an empirical bias exists in Di Tella, et al.'s, data, it seems not to be one of cynicism, but rather one of optimism and positivity that arises only among those with little ability to take from their opponent.

We next show that, theoretically, this positivity bias can be explained by a simple model that serves as a stylized representation of the games in both ours and Di Tella, et al.'s, experiments. In the model, Ann derives utility from her own and Bob's payoffs and this utility is increasing in Bob's kindness. When Ann has the opportunity to take from Bob, she has an incentive to reduce her belief regarding Bob's kindness; this diminishes the loss in utility she experiences by taking from him. However, in this setting individuals also have an incentive to form another kind of motivated belief—to convince themselves that the other player is kind and deserves any payoff she receives. The net result of these two opposing tendencies is an absolute bias in the direction of positivity, which is consistent with a general tendency for distorted beliefs to lie in the direction of positivity and optimism, rather than the opposite (Bénabou and Tirole, 2016). This simple theoretical analysis provides a basis for two phenomena we observe in our first experiment and in Di Tella, et al.'s, data. First, in absolute terms, biases about others' actions will lie in the direction of positivity rather than cynicism.[4] Second, consistent with Di Tella, et al.'s, interpretation of their findings, the relative positivity of Ann's beliefs about Bob's behavior will be lower as Ann has a greater opportunity to take money from Bob. Thus, viewed jointly, these predictions suggest that, at least in many settings, strategic cynicism may be a relative rather than an absolute phenomenon.

The above analysis yields a straightforward interpretation for the absence of absolute strategic cynicism in our first experiment and the presence of relative strategic cynicism in

[4] This prediction also arises under the model that Di Tella, et al., use to motivate their experiment, though their analysis does not investigate this property.

the study by Di Tella, et al. However, existing empirical support for the above two predictions involves comparisons across studies, in which changing populations or incidental factors may yield varying results. Therefore, in a third step, we test the two predictions in a novel experiment. We conduct a replication of Di Tella, et al.'s, study, but also elicit the beliefs of neutral observers regarding the behavior of subjects in the role of Bob.[5]

In this new experiment, we replicate Di Tella, et al.'s, main finding—a comparative static result that individuals with a greater opportunity to take from their counterparts hold less positive beliefs about these opponents. This replication itself is noteworthy, as we have a substantially larger sample size and find qualitatively similar findings in a different population, in Switzerland rather than Argentina, in a society that differs in general levels of corruption, trust and trustworthiness. However, we also once more document a lack of strategic cynicism in absolute terms. The beliefs of individuals in the role of Ann with a strong incentive to engage in strategic cynicism are no more cynical about Bob's behavior than either the empirical frequency of actual choices or the beliefs of neutral third parties without any incentive to adopt a negative view of Bob's likely actions.

Our results should not be interpreted as questioning Di Tella, et al.'s, findings. In fact, we provide a direct replication of their main result of relative strategic cynicism. However, we additionally provide clear evidence—across both of our studies and in Di Tella, et al.'s, original data—that there is no strategic cynicism in absolute terms. Instead, we find that strategic beliefs are positively biased. Our contribution thus expands our understanding of the psychological forces behind self-serving belief manipulation, by noting that strategic cynicism may compete with a tendency towards positivity in determining individuals' beliefs. Such a tendency towards positivity is consistent with overwhelming evidence of a general "positivity illusion" (Taylor and Brown, 1988) from psychological studies: people hold overoptimistic beliefs about future life events (Weinstein 1980, 1989), are too optimistic about the degree of personal control (Langer, 1975), hold too positive perceptions of themselves (Svenson, 1981; Quattrone and Tversky, 1984), engage in wishful thinking (Irwin, 1953), and hold beliefs that the world is just (Lerner, 1980).[6] This positivity bias can also explain why, in contrast to our study, many other studies found strong evidence for

[5] Di Tella, et al., argue that one of their experimental treatments provides an estimate of unbiased beliefs. However, as we discuss in detail below (see footnote 17), there are a few reasons why these estimates are unlikely to correspond to the unbiased beliefs of subjects in their main experiment.

[6] In economics, Haisley and Weber (2010) document a tendency to believe that the impacts of one's choices on others are more positive than they actually are, while Andreoni and Sanchez (2014) find that subjects are too optimistic about other players' trust and trustworthiness compared to actual behavior.

motivated reasoning (in non-strategic settings). We therefore contribute to a better understanding of the specific contexts, in which we should expect biased beliefs to arise.[7]

The next section provides a detailed description of our first study using the pre-emptive taking game. In Section 3, we discuss the study by Di Tella, et al., show that their findings do not provide evidence of strategic cynicism in absolute terms and present a stylized model that can provide an interpretation of behavior in both experiments. Section 4 presents our second experiment, intended to test this model more directly and reconcile the earlier results. Finally, Section 5 concludes.

2. Study 1: An experimental test of strategic cynicism

We first introduce the pre-emptive taking game, which allows us to test for absolute bias in strategic beliefs. Then, we discuss the experimental implementation of the game and present our results.

2.1 The pre-emptive taking game

There are two players, Ann and Bob. Both players start with an endowment of 10. They play a sequential game. In Stage 1, Ann decides how much to take from Bob's initial endowment. She can take any amount, $a \in \{0, 2, 4, \ldots, 10\}$. After Stage 1, Ann's wealth equals $10 + a$, while Bob's equals $10 - a$.

In Stage 2, after observing a, Bob decides how much to take from Ann's current endowment, b, once again in increments of two. The amount that Bob can take is constrained by Bob's remaining wealth. Specifically, in order to take b units from Ann, Bob has to spend $0.5b$ from his remaining wealth and cannot spend more than the amount he has at the beginning of Stage 2. Furthermore, Bob cannot take more than Ann's wealth at the beginning of Stage 2. Thus, Bob's ability to take is given by, $b \in \{0, \ldots, \bar{b}\}$, where $\bar{b} = \min(2(10 - a), 10 + a)$. Hence, in the case in which Ann took everything in Stage 1 (i.e., $a = 10$), Bob cannot take anything in Stage 2.

After Stage 2, the game concludes. The two players' payoffs are determined as follows:

$$\pi_A = 10 + a - b$$

[7] Other studies that demonstrate limits in the extent to which motivated reasoning and justifications facilitate egoistic behavior are van der Weele, Kulisa, Kosfeld and Friebel (2014) who find that people do not use "moral wiggle room" (see Dana, Weber and Kuang, 2007) in the context of reciprocity and Bartling and Özdemir (2017) who find that people do not employ the "replacement logic" ("if I don't do it, someone else will") in contexts with a strong social norm.

$$\pi_B = 10 - a + b - 0.5b$$

As an example, suppose Ann decides to take 6 in Stage 1 such that $a = 6$. At the beginning of Stage 2, Ann has 16 and Bob has 4. In this case, in Stage 2, Bob can spend up to 4 to take up to 8 from Ann; doing so would leave both Ann and Bob with final payoffs of 8. Under standard egoistic preferences, in the unique subgame-perfect Nash equilibrium to the game both Ann and Bob take as much as they can, i.e., $a = 10$ and, consequently, $b = \bar{b} = 0$.

The key feature in the pre-emptive taking game is that Bob's ability to take from Ann is limited by how much he has left at the end of Stage 1. Ann can thus protect herself from Bob's potentially egoistic behavior by taking all his tokens. Therefore, suppose Ann wants to obtain as high a payoff as possible, but also feels obligated to be fair to Bob in the case he does not intend to take from her. In such a case, Ann may justify taking by convincing herself that Bob intends to act greedily—i.e., by engaging in strategic cynicism. The critical measure of strategic cynicism in studying this game is thus Ann's beliefs about Bob's behavior. In particular, eliciting these beliefs and comparing them to neutral and objective standards—the actual amount of taking by Bob and neutral observers' beliefs about Bob's taking—allows us to test whether they exhibit a systematic bias toward negativity.

2.2 Experimental design

At the beginning of each session, participants are randomly assigned to one of three roles: Player A (Ann), Player B (Bob) and Player C (neutral observer). Subjects are informed of their own role. Next, all subjects receive the same set of instructions. The instructions describe all decisions made by Player A, Player B and the neutral observer in detail and subjects are provided with a detailed table showing all the possible combinations of payoffs resulting from the two strategic players' actions.[8] After hearing the instructions read aloud, all participants answer questions about the decisions available to Players A and B and the consequences of these decisions.

For the pre-emptive taking game, Players A and B each start with an initial endowment of 10 chips, with each chip worth CHF 2 (≈ $2). In each pair, Player A selects how much to take from Player B (a). Player B's choices are elicited using the strategy method—Player B selects an amount to take (b_a) for every possible choice made by Player A. After both Player A and Player B have made their decisions, but before they learn about the payoffs, we elicit their beliefs concerning their counterpart's behavior. Player B guesses

[8] The instructions are available in the Appendix.

which value of a Player A selected ($\hat{a}^B$). Player A guesses a value of b for every possible value of a, or $\hat{b}_a^A$; at the end of the experiment one value of a is randomly selected to count for Player A's guess. Each subject earns an additional CHF 4 if they accurately guess the choice made by their opponent.[9]

Individuals in the role of Player C are not matched with any pair and are not directly affected by the choices made by any specific Player A or B. Hence, they act as neutral participants, who have no incentive to bias their beliefs about other participants' actions. For our purposes, they provide a measure of unbiased beliefs about the actions of Player As and Bs. Specifically, each neutral observer guesses the choice of a randomly selected Player A ($\hat{a}^C$) and the conditional choices of a randomly selected Player B ($\hat{b}_a^C$). Similarly to the other participants, each Player C gains CHF 4 for correctly guessing the behavior of a Player A and CHF 4 for correctly guessing one randomly selected option for a Player B.

After making all choices, participants are informed about their payoffs. They then answer several socio-demographic questions before they are paid in private.

We conducted seven sessions with between 30 and 36 participants, resulting in a total of 240 participants, 80 in each role.[10] All sessions took place at the Decision Sciences Lab (DeSciL) at the Federal Institute of Technology (ETH) in Zurich in 2015. Participants were recruited using hroot (Bock, Baetge and Nicklisch, 2014) from the joint subject pool of the University of Zurich and the ETH. The experiment was implemented using z-Tree (Fischbacher, 2007).

2.3 Results

On average, Player A took 8.0 tokens (std. dev. = 3.6) from Player B, with 50 of 80, or 62.5 percent, taking the full amount, $a = 10$. Figure A1 in the Appendix provides the full distribution of amounts taken.

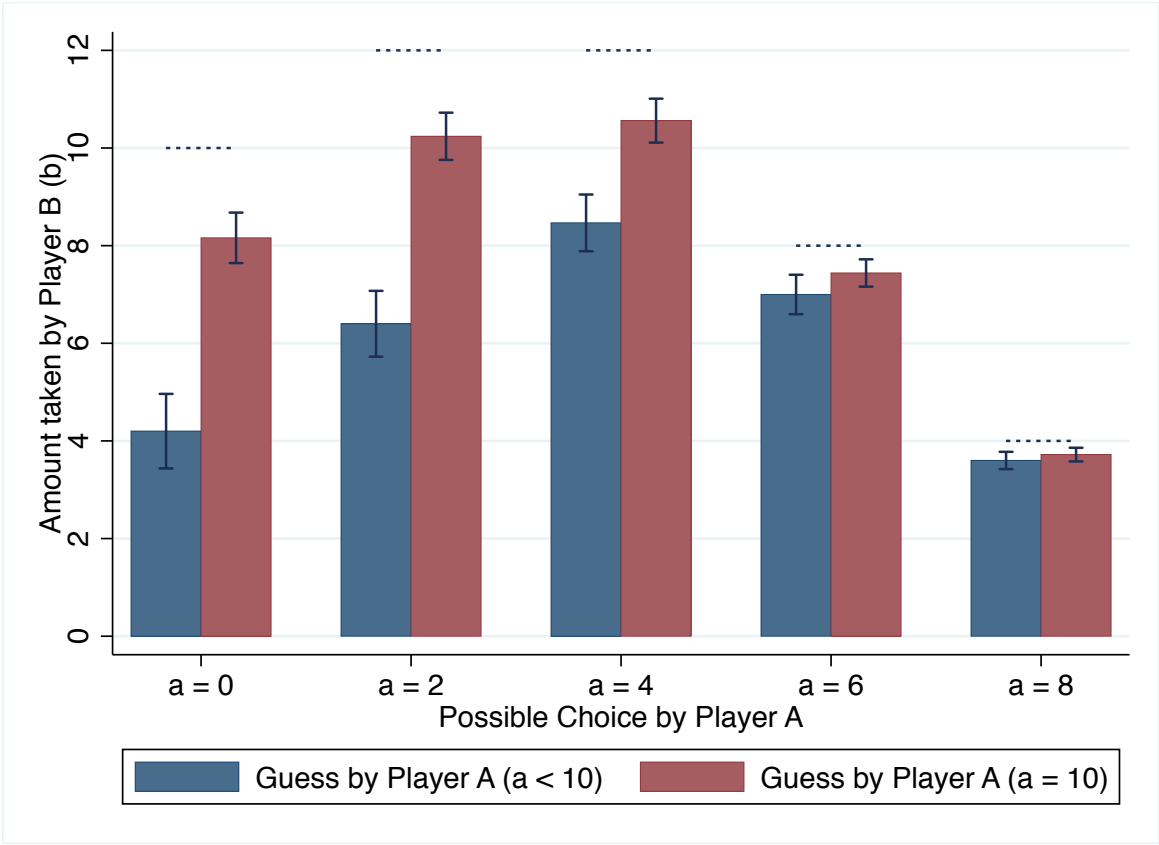
Taking by Player A is related to beliefs about how much Player B will take. Figure 1 shows the average belief of Player A regarding how much Player B will take, in response to every possible action by Player A. The figure presents these mean beliefs separately for those who took less than 10 ("a < 10") and those who took everything ("a = 10"). Those subjects in

[9] The amount of CHF 4 as an incentive for accurate guesses was the same for all sessions, except of the first session. In this session, the incentive for accurate guesses was CHF 2. We raised the incentive subsequently to provide subjects with more earnings opportunity. We find no differences in accuracy of guesses due to different incentives.

[10] We conducted two waves: the first four sessions were in Wave 1 while the remaining three sessions were in Wave 2. The second wave included elements intended to better ensure comprehension. We pool the data, as there is no difference in behavior between the two waves. The appendix provides instructions for both waves.

the role of Player A who took everything hold more cynical beliefs about Player B. For instance, for the hypothetical case where Player A takes nothing ($a = 0$), those who actually took all 10 have mean beliefs that are much more cynical (8.16) than those who took less than 10 (4.2), and this difference is highly statistically significant ($t_{78} = 4.44$, $p < 0.001$). Comparisons of mean beliefs for the cases in which Player A takes $a = 2$ or $a = 4$ similarly reveal differential cynicism between those who took 10 and those who took less (respectively, $t_{78} = 4.72$, $p < 0.001$ and $t_{78} = 2.85$, $p < 0.01$). This positive relationship between taking by Player A and negative beliefs about Player B's behavior is consistent with strategic cynicism but does not demonstrate it. Indeed, the more straightforward interpretation is that subjects in the role of Player A might simply be responding to their beliefs—taking more preemptively if they fear that B will also take more.[11]

Figure 1: Player A beliefs about Player B's actions by Player A type



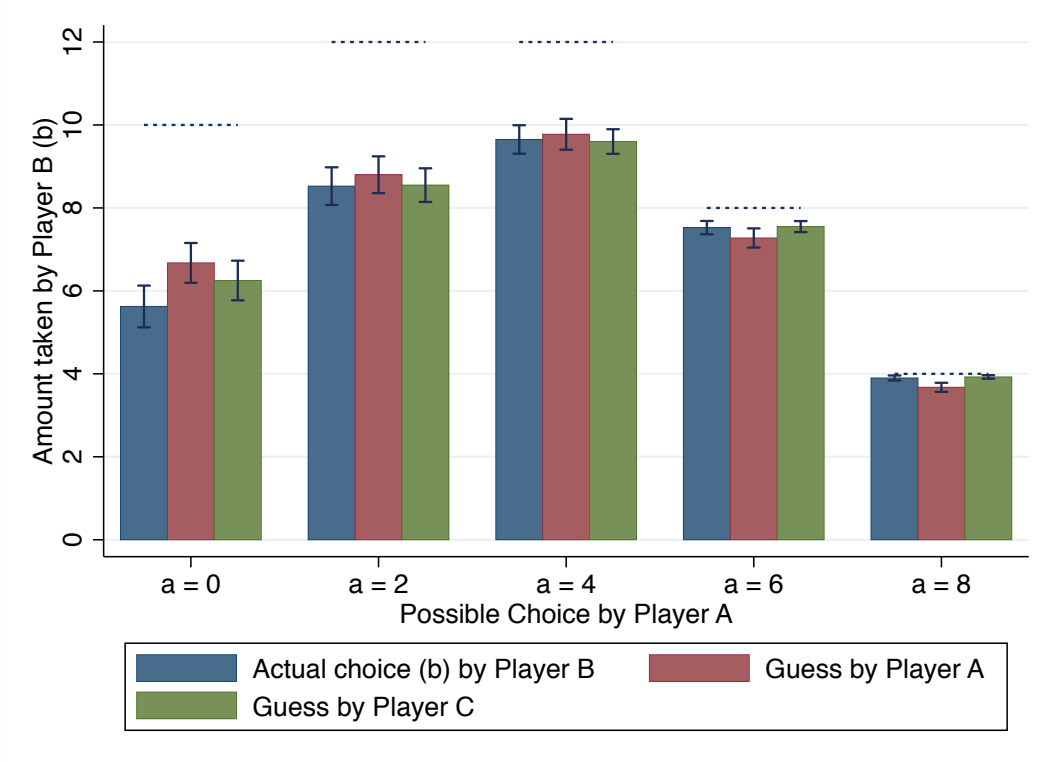
Note: The figure displays mean predictions by A of how much B will take, conditional on how much they took themselves. The category "a<10" represents Player As who took less than 10 from Player B, while the category "a=10" represents Player As who took the maximum possible from their counterparts. Dotted lines indicate the maximum possible amount B could take. Bars indicate standard errors.

[11] This is in line with results in Fischbacher and Gächter (2010) who find, using a public good game, a positive relationship between participants' own contributions and their beliefs about others' contributions.

To investigate strategic cynicism, our main focus, we next compare the beliefs of subjects in the role of Player A with two standards corresponding to unbiased beliefs. Strategic cynicism would imply that the beliefs of Player A about B's action predict more taking than the actual amount taken by Player B ($\hat{b}_a^A > b_a$) and than the corresponding predictions by Player C ($\hat{b}_a^A > \hat{b}_a^C$). Figure 2 shows, for each possible action by Player A, how much Player B actually took on average and the corresponding mean beliefs by subjects in the roles of Player A and Player C.

Looking first at the actual behavior of subjects in the role of Player B, we observe that the amount they take depends on Player A's choice. For instance, when Player A takes nothing, then the average amount taken, $b_{a=0}$, equals 5.63, even though Player B could take anywhere between 0 and 10 tokens. As Player A takes more, Player B also seizes a larger proportion of the available tokens. For instance, when Player A takes either 2 or 4 tokens, meaning that Player B can take anywhere between 0 and 12, then $b_{a=2}$ = 8.53 and $b_{a=4}$ = 9.65, on average. Finally, when A takes most of B's endowment, B takes, on average, very close to the maximum possible amounts ($b_{a=6}$ = 7.53 and $b_{a=8}$ = 3.90).[12]

Figure 2: Actual conditional taking by B and corresponding predictions by A and C



Note: Actual choices by Player Bs and predicted choices about Player B's behavior by Player As and the neutral observers (Player Cs), respectively. Dotted lines indicate the maximum possible amount B could take. Bars indicate standard errors.

[12] Figure A2 in the Appendix provides the distributions of taking by B, for every possible amount taken by A.

Perhaps the most striking finding in Figure 2 is how little evidence we find of strategic cynicism. For every possible amount taken by Player A, the beliefs of Player A and Player C regarding Player B's choice are very close to each other and to the actual behavior of Player B. Table 1 presents statistical tests of the relationships between Player A's beliefs and the actual choices by Player B and the beliefs of Player C. There are no statistically significant differences between the beliefs of Player A ($\hat{b}_a^A$) and the actual behavior of Player B (b_a) or the beliefs of Player C ($\hat{b}_a^C$) for any amount taken by A between 0 and 6. In the case where A takes 8, the differences are at least marginally statistically significant, but in these comparisons A *underestimates* B's taking, both relative to the actual amount and to the beliefs provided by C.[13]

Table 1. Statistical tests of strategic cynicism

	Mean taking by B (b_a)	Mean guess by A ($\hat{b}_a^A$)	Mean guess by C ($\hat{b}_a^C$)	$\hat{b}_a^A$ vs. b_a	$\hat{b}_a^A$ vs. $\hat{b}_a^C$
$a = 0$	5.625 (0.503)	6.675 (0.480)	6.250 (0.478)	$t_{158} = 1.510$ p = 0.133	$t_{158} = 0.678$ p = 0.531
$a = 2$	8.525 (0.454)	8.800 (0.444)	8.550 (0.406)	$t_{158} = 0.433$ p = 0.666	$t_{158} = 0.416$ p = 0.678
$a = 4$	9.650 (0.343)	9.775 (0.371)	9.600 (0.296)	$t_{158} = 0.247$ p = 0.805	$t_{158} = 0.368$ p = 0.713
$a = 6$	7.525 (0.160)	7.275 (0.231)	7.550 (0.133)	$t_{158} = 0.889$ p = 0.375	$t_{158} = 1.031$ p = 0.304
$a = 8$	3.900 (0.061)	3.675 (0.109)	3.925 (0.043)	$t_{158} = 1.800$ p = 0.074	$t_{158} = 2.130$ p = 0.035

Standard errors in parentheses

2.4 Discussion

To summarize, we find a positive relationship between cynicism about one's opponent and the number of tokens taken by subjects in the role of Player A. That is, subjects who hold cynical beliefs about their opponents take more from them. However, our data

[13] The statistical tests in Table 1 are t-tests. Non-parametric Wilcoxon rank-sum tests yield very similar results; all p-values for $a = 0$ through $a = 6$ are greater than 0.188 and the p-values for $a = 8$ are 0.071 ($\hat{b}_a^A$ vs. b_a) and 0.066 ($\hat{b}_a^A$ vs. $\hat{b}_a^C$).

reveal very little evidence of self-serving belief manipulation by subjects in the role of Player A, neither relative to objective behavioral standards nor to the beliefs of unbiased observers.[14]

While our results rule out significant levels of strategic cynicism in our data, they stand in contrast to Di Tella, et al. (2015), who conclude from their experimental evidence that people engage in self-serving manipulation of their strategic beliefs. We next describe this evidence and attempt to reconcile our seemingly conflicting results.

3. Reconciling our results with Di Tella, et al. (2015)

Di Tella, et al. (2015), study strategic cynicism using a "corruption game." In contrast with our findings, they conclude that decision makers bias their beliefs about a counterpart's egoism in response to incentives. In this section, we first describe their study and findings in detail and then offer evidence that our two sets of results are similar in that neither yields evidence of an *absolute bias* in individuals' beliefs.

3.1 Di Tella, et al.'s, corruption game

The study by Di Tella, et al., is based on the idea that a greater opportunity to act egoistically at the expense of a counterpart creates stronger incentives to engage in strategic cynicism. Hence, their main prediction is that those with greater opportunities to act egoistically should end up with a more pessimistic belief about the counterpart's kindness.

Di Tella, et al., test this comparative-static prediction in a "corruption game."[15] In the game, an "Allocator" and a "Seller" each start with 10 tokens. The Allocator decides how to redistribute the combined 20 tokens between herself and the Seller. Simultaneously, the Seller sets the "price" at which all the tokens are sold to the experimenter. He can either choose a price of 1.50 Argentine pesos ($) or a price of $0.50. If the Seller chooses the latter, he additionally receives a fixed side payment of $5. Setting a lower price while taking the side payment is labeled as a "corrupt" act by the Seller, akin to accepting a bribe.

In addition to deciding upon the distribution of tokens, the Allocator also provides beliefs ($\hat{p}$) about the likelihood that her paired Seller takes the corrupt action and about the share of Sellers in the experimental session that does so. The Allocator receives $5 for each

[14] The beliefs of Player Bs and Cs about A's actions are similarly unbiased (see Figure A3 in the Appendix). Recall that the mean amount taken by A is 8.0 (std. dev. = 0.358). Player Bs and C report mean beliefs that are slightly higher ($\hat{a}^B$ = 8.35 (0.280), $\hat{a}^C$ = 8.53 (0.258)) but these differences are not statistically significant—all comparisons using t-tests or rank-sum tests yield $p > 0.23$.

[15] Di Tella, et al., present the results of two different corruption games. We describe the game used in their preferred study, the modified corruption game. The formal games only differ in their payoffs.

correct guess. These beliefs—the key measure in Di Tella, et al.'s, experiment—provide estimates of the (possibly biased) beliefs that the Allocator has about Sellers' behavior.

Di Tella, et al., identify strategic cynicism with a treatment distinction that varies constraints on Allocators' ability to redistribute tokens. In the "Able = 2" treatment, the Allocator can move up to two tokens; that is, she can implement one of the following five payoff distributions: (8, 12), (9, 11), (10, 10), (11, 9), (12, 8). In the "Able = 8" treatment, the Allocator can move up to eight tokens, meaning that the allocations, (2, 18), (3, 17), …, (17, 3), (18, 2), are all possible. Hence, the treatment manipulation endows some Allocators with the ability to appropriate up to eight of the Sellers' tokens and other Allocators with the ability to appropriate only up to two tokens. Importantly, however, a Seller is not informed of whether his paired Allocator can move two or eight tokens, meaning that Allocators, who are aware of Sellers' ignorance regarding the treatment, should form beliefs about Sellers' behavior that are independent of the treatment.[16] Instead, Di Tella, et al., argue that the treatment manipulation affects the value of engaging in self-deception, as "allocators who can take more tokens from the seller (i.e., Able = 8 instead of Able = 2) have more incentives to convince themselves that the seller is unkind" (p. 3417), that is, $\hat{p}_{Able=2} < \hat{p}_{Able=8}$.

Di Tella, et al.'s, experimental results are consistent with this prediction. Individuals who have a greater ability to take from their counterpart take more and, more importantly, reveal more pessimistic beliefs about their counterpart's corruption. This appears to contrast with the findings of our first experiment, which reveal no strategic cynicism.

3.2 Relative or absolute cynicism?

To reconcile the apparent discrepancy, first note that our first study sought to identify strategic cynicism through the observation that individuals bias their beliefs about a counterpart's actions relative to the objective reality or to the beliefs of unbiased observers. Specifically, suppose there is some measure of an opponent's (expected) kindness, d—where unkindness, or $1 - d$, corresponds to b in the pre-emptive taking game or p in the corruption game. If $\hat{d}$ represents a decision maker's beliefs about the opponent's kindness, then our strategic cynicism hypothesis is that a decision maker who can take from the opponent will perceive the opponent to be less kind than he actually is, or $\hat{d} < d$. In the case of Di Tella et al.'s experiment, for instance, this corresponds to the belief that p is higher than it actually is

[16] This design leaves open the possibility that Allocators' differential beliefs are the result of the "curse of knowledge" or information projection (Camerer, Loewenstein & Weber, 1989; Madarasz, 2012), whereby decision makers find it difficult to ignore their private information when guessing others' beliefs.

(see footnote 3). Instead, our first study finds that $\hat{d} \approx d$, both when we measure d using empirical behavior as the benchmark or the beliefs of unbiased observers.

Di Tella, et al.'s, experiment and findings, however, demonstrate something different. Specifically, they test whether a decision maker with a greater incentive to take from the opponent will perceive the opponent to be less kind than will a decision maker with a reduced incentive to take. That is, if we let $\hat{d}'$ represent the beliefs of a decision maker with a restricted taking opportunity—as with Allocators in the Able = 2 condition—then Di Tella, et al.'s, findings demonstrate that those who are constrained to take less adopt relatively more positive beliefs of their opponent's kindness, $\hat{d}' > \hat{d}$. Indeed, this relative comparison is also the basis of the main theoretical proposition with which they motivate their study.

The discrepancy in our findings is straightforward to reconcile if one recognizes that a relative bias and an absolute bias may not coincide. That is, if subjects in Di Tella, et al.'s, study do not exhibit an absolute bias in the direction predicted by strategic cynicism—that is, if $\hat{d}' > \hat{d} \geq d$—then the two sets of results are entirely consistent.

Di Tella, et al., do not explicitly collect a measure of the unbiased beliefs of neutral observers.[17] However, we can compare Allocators' guesses to the empirical frequency of Sellers' corruption. In Di Tella, et al.'s, first experiment, the actual proportion of Sellers who chose the corrupt option was 75 percent. Using the more precise measure of Allocators' estimates, and the only one that was incentivized, we see that those Allocators with a greater ability to take money (Able = 8) provided estimates (69 percent) that were fairly close to the empirical benchmark and, if anything, *underestimated* Sellers' corruption. On the other hand, the Allocators who had a reduced ability to take (Able = 2) provided estimates (49 percent) that were much farther from the true percentage. Similarly, in the second experiment, the frequency of corrupt behavior by Sellers was 66 percent. The estimates provided by those who could take more (Able = 8) tended to exhibit very little bias (64 percent), while the estimates from those who could take less (Able = 2) were again biased in the direction of believing too little corruption on the part of Sellers (48 percent).

[17] Di Tella, et al., argue that an additional treatment in which Allocators are forced to take a pre-specified amount from the Seller, and in which the mean estimate is 56 percent, provides "a rough estimate of what the average [estimate] would have been in the Modified Game if Allocators had not incurred in self-deception." However, these are not the estimates of unbiased observers, but of individuals interacting with the counterpart. In the simple model we develop below, such individuals have an incentive to view the opponent kindly. In addition, eliminating the Allocator's choice altogether substantively changes the game—e.g., Sellers now confront a unilateral decision problem rather than a strategic game. Therefore, the beliefs of the Allocators in this game can only very cautiously be interpreted as corresponding to Allocators' (unbiased) beliefs in the corruption game. Finally, there are very few observations in this treatment (15 if one uses the same rules for excluding subjects as in other treatments—see footnote 25 in Di Tella, et al., 2015).

Our observations—based on our data and that of Di Tella, et al.—suggest that, to the extent a bias exists, it is one of positivity and the belief that opponents will be kinder than they actually are. While Di Tella's, et al.'s, evidence points to greater relative strategic cynicism on the part of those with more opportunity to take—i.e., $\hat{d}' > \hat{d}$—there is very little evidence of strategic cynicism on an absolute level—instead, it appears that $\hat{d}' > \hat{d} \geq d$. We next demonstrate how a very simple model can provide an interpretation for these patterns.

3.3 A simple model of strategic cynicism

In this section, we introduce a simple model that can account for the above patterns. We do not attempt to derive a general model of belief formation or self-deception. Instead, we study a simple and highly stylized representation of a decision in a non-strategic context, where an individual decides on a wealth allocation between herself and a counterpart and cares about this counterpart's perceived kindness or unkindness. It shares many features with the model used by Di Tella, et al., to motivate their experiment—indeed, the main predictions of our analysis can also be generated using their model.[18] We acknowledge that alternative modeling approaches may yield different predictions; however, we present this model merely as an example of the kind of model that can provide an interpretation for the above patterns we observe in our first experiment and in the study by Di Tella, et al., and that we can further test in a novel experiment.

Ann decides how to split an amount of money, normalized to 1, with Bob. Ann can take at most $K \in (0,1]$ for herself. Bob has one of two possible types, which represent the extent to which he is a "good" or "bad" person and is therefore perceived by Ann to deserve greater or less wealth: with probability, $p \in (0,1)$, Bob is a low-deservingness type (L) and with probability, $1-p$, Bob is a high-deservingness type (H). Ann is altruistic, but she cares less for the welfare of the low type. Specifically, she puts weight $d_L > 0$ on the low type's payoff and weight $d_H > d_L$ on the payoff of the high type. To incorporate motivated reasoning and self-deception, Ann can bias her belief, $\hat{p}$, about the share of low types. However, this incurs a psychological cost, $C(|p-\hat{p}|)$.[19] In addition, as in our experiments and those of Di Tella, et al., Ann is incentivized to hold unbiased beliefs by a monetary

[18] Our model differs in important points (e.g., we formally incorporate the restriction K). Di Tella, et al., discuss a result related to Proposition 2, but no result related to Proposition 1.

[19] We capture motivated reasoning similarly to Rabin (1994) and Konow (2000). For other examples of models with motivated beliefs, see Bénabou and Tirole (2006, 2011), Bénabou (2013), Brunnermeier and Parker (2008), and Bodner and Prelec (2002, 2003).

payoff function $P(|p-\hat{p}|)$. Since Ann does not know Bob's actual type, the weight she assigns to Bob's payoff equals its expected value, $E_{\hat{p}}(d) = \hat{p}d_L + (1-\hat{p})d_H$.

Ann's behavior is then captured by the following maximization problem:

$$\max_{x\in[0,K],\hat{p}\in[0,1]} \mathrm{U}(x,\hat{p}) = v(x) + E_{\hat{p}}(d)v(1-x) + P(|p-\hat{p}|) - C(|p-\hat{p}|), \quad (1)$$

where x represents the share that Ann takes for herself. The function $v\colon [0,1] \to \mathbb{R}_{>0}$ is a C^1 function with $v' > 0$, $P\colon [0,1] \to \mathbb{R}_{\geq 0}$ is a C^2 function with $P' < 0$ and $P'' \leq 0$, and $C\colon [0,1] \to \mathbb{R}_{\geq 0}$ is a C^2 function with $C' > 0$ and $C'' > 0$.[20]

This model generates two predictions that are consistent with the above empirical observations (all proofs are in Appendix B). First, Proposition 1 shows that Ann is, if anything, too positive regarding Bob's deservingness.

Proposition 1: For any $(x,\hat{p})$ *in* $argmax_{x\in[0,K],\hat{p}\in[0,1]}\ \mathrm{U}(x,\hat{p})$, $\hat{p} \leq p$.

This result is due to the fact that her utility increases in Bob's deservingness; that is, Ann prefers to think of Bob as being the more deserving type. Thus, in absolute terms, Ann's bias will always be to think of Bob as kinder than he actually is.

Proposition 2 states that, in relative terms, Ann will be more cynical—or less positive—when she has the opportunity to take more from Bob.

Proposition 2: Take K, K' *in* $(0,1]$ *with* $K' < K$ *and suppose that there is a unique solution to (1) for both* K *and* K'*, then* $\hat{p}' \leq \hat{p}$.

Thus, Proposition 2 provides a basis for the differences in relative beliefs between subjects who could take varying amounts from their opponent in Di Tella, et al.'s, experiment. In summary, this model predicts a (motivated) bias in belief formation that provides a basis for the absence of absolute levels of cynicism: as Ann shares a positive amount with Bob, she likes to think of him as a deserving type and this motivation is stronger as she is constrained to take less money from him.

Next, we illustrate the choice problem for a neutral observer with no stake in the outcome of game. This individual reports his beliefs, $\hat{p}^N$, about the deservingness of a random person in the role of Bob and faces incentives for accuracy: he receives a payoff

[20] Note that there exists a solution to (1) as $\mathrm{U}(x,\hat{p})$ is continuous and the feasible set is compact.

$P(|p - \hat{p}^N|)$. The observer can also adopt biased beliefs about p, but faces the cost, $C(|p - \hat{p}^N|)$, for doing so. The observer thus solves the following maximization problem:

$$\max_{\hat{p}^N \in [0,1]} U(\hat{p}^N) = P(|p - \hat{p}^N|) - C(|p - \hat{p}^N|) \quad (2)$$

where again $C' > 0$ and $P' < 0$. The observer maximizes his utility by having accurate beliefs:

Proposition 3: $\hat{p}^N = p$ *is the unique solution to* $\max_{\hat{p}^N \in [0,1]} U(\hat{p}^N)$.

Therefore, neutral observers report beliefs that correspond to the unbiased beliefs p.[21]

Note that this simple model can account for the pattern observed in the above laboratory results. Our results in Study 1 are consistent with Propositions 2 and 3: $\hat{p} \leq p = \hat{p}^N$. In the case of Di Tella, et al.'s, experiment, if we allow Able = 2 to correspond to K' and Able = 8 to correspond to K, the patterns in the data are consistent with both Propositions 1 and 2: $\hat{p}' \leq \hat{p} \leq p$. We next test these predictions jointly—specifically, that $\hat{p}' \leq \hat{p} \leq \hat{p}^N = p$—in a novel experiment.

4. Study 2: Jointly testing absolute and relative strategic cynicism

As we state above, the data from our first experiment and that of Di Tella, et al., provide, separately, support for all three of the above propositions. However, since we developed the model as a way to account for these observations, this is not particularly surprising. We next report a novel study that jointly tests all three predictions. The new experiment replicates Di Tella, et al.'s, experiment and further elicits the beliefs of neutral observers.

4.1 Experimental design

We began with a replication of Di Tella, et al.'s, corruption game, conducted in Switzerland. We used their instructions and replaced the monetary payoff of 1 Argentine peso with *1.20* Swiss Francs (CHF).[22] This meant, for instance, that the payment for each correct guess by an Allocator was *CHF 6*. In addition, we paid a participation fee of *CHF*

[21] See Konow (2000) for a very similar result, in the case of "Benevolent Dictators."
[22] Note that we substantially increased real incentives; in 2016, CHF 1 corresponded to 7.48 Argentine pesos (PPP adjusted; OECD, 2018).

15.[23] We made two further substantive changes to bring their classroom experiment into a laboratory setting. First, while their experiment was fully paper based, we implemented it via computers, using the z-Tree program (Fischbacher, 2007). Second, we used a slightly different procedure to guarantee participants' anonymity.[24] We conducted nine sessions, each with between 22 and 24 participants, resulting in a total of 212 participants (106 Allocators and 106 Sellers).

We also conducted an additional variant of the experiment—which we refer to as the "neutral" treatment—to elicit the unbiased beliefs of neutral observers. In these sessions, participants received "instructions provided to a participant in a previous experiment." Specifically, each participant saw either the instructions given to an *Able = 2* or an *Able = 8* Allocator, determined at random. Since the Allocators' instructions include the instructions given to Sellers, participants also read the Sellers' instructions and had knowledge of the entire game. We made explicit to participants that, first, these were not their instructions and that, second, at the end of the experiment they could earn money by providing accurate guesses about something that happened in the previous experiment. Therefore, participants had no incentives to engage in self-deception, but still had incentives to closely attend to the instructions and understand the corruption game.

After reading the instructions, participants in the neutral treatment first had to answer the same comprehension questions as in the design of Di Tella, et al., and in our replication, to make sure that they understood the game. Subsequently, they made two guesses identical to those made by Allocators in Di Tella, et al.'s, design and in our replication: they guessed the choice made by a randomly chosen Seller in the previous experiment and they guessed what percentage of Sellers in a previous session chose the corrupt option. As in the replication, they received *CHF 6* for correct guesses, as well as a *CHF 15* participation fee. We conducted two such neutral sessions, with a total of 55 participants.

All sessions took place at the Decision Sciences Laboratory at the Federal Institute of Technology (ETH) in Zurich, in 2016. Participants were recruited using hroot (Bock, Baetge

[23] Our first session paid only a participation fee of CHF 10. We adjusted this payment upward to reflect the longer duration of the study than we originally expected.

[24] At the beginning of each session, one participant was randomly selected to be the "monitor." The remaining participants each received a random ID number hidden in an envelope, so that the experimenter could not match the ID to the participant. Subjects entered their ID numbers in their respective computer terminals. At the end of the study, we placed the amount of money earned by each participant in an envelope labeled only with the anonymous ID number and placed all envelopes on a table that participants passed on their way out of the laboratory. The monitor, who did not know the amount contained in any of the envelopes, controlled that each participant took only the correct envelope.

and Nicklisch, 2014) from the joint subject pool of the University of Zurich and the ETH. All instructions for the replication and the neutral treatment are in the Appendix.

Table 2. Allocator behavior in Di Tella, et al., and our replication

	Modified Game (Di Tella, et al.)			*Replication*		
	Able = 2	*Able = 8*	*p-value*	*Able = 2*	*Able = 8*	*p-value*
Tokens Taken	1.35 (0.16)	6.59 (0.33)	<0.01	0.98 (0.14)	4.79 (0.44)	<0.01
Is Corrupt	0.48 (0.09)	0.85 (0.06)	<0.01	0.19 (0.05)	0.49 (0.07)	<0.01
%-Corrupt	0.48 (0.04)	0.64 (0.03)	<0.01	0.33 (0.04)	0.47 (0.04)	<0.01
N	31	34		53	53	

Notes: Standard errors in parentheses. P-value: t-test of the null hypothesis that the means under Able = 2 and Able = 8 are equal.

4.2 Results

Table 2 compares the behavior and guesses of Allocators in Di Tella, et al.'s, experiment and in our replication.[25] The mean tokens taken (*Tokens Taken*) by the Allocator is slightly lower in our replication. More importantly, the share of Allocators who think that their paired Seller chooses the "corrupt" side payment (*Is Corrupt*) differs significantly between the *Able = 2* and *Able = 8* treatment groups in both the original experiment and our replication. The same holds for the average stated belief of Allocators regarding the share of Sellers who choose the side payment (*%-Corrupt*). Thus, despite the differences between the two studies—e.g., lab vs. field, Switzerland vs. Argentina—we replicate Di Tella, et al.'s, findings of relative strategic cynicism ($\hat{p}' \leq \hat{p}$).[26]

[25] Following Di Tella, et al., we conduct randomization tests with respect to demographic measures (gender, age and socioeconomic class). We find no significant differences between the Able = 2, Able = 8 and Neutral treatments.

[26] The statistical tests in Table 2 are t-tests. Non-parametric Wilcoxon rank-sum tests yield very similar results; p-values for Tokens Taken, Is Corrupt and %-Corrupt are smaller than 0.01.

Table 3. Allocator and neutral observer beliefs

	Replication		*Neutral Treatment*		
	Able = 2	*Able = 8*	*Neutral*	*p-value vs. Able = 2*	*p-value vs. Able = 8*
Is Corrupt	0.19 (0.05)	0.49 (0.07)	0.44 (0.07)	0.005	0.576
%-Corrupt	0.33 (0.04)	0.47 (0.04)	0.46 (0.04)	0.019	0.743
N	53	53	55		

Notes: Standard errors in parentheses. P-value: t-test of the null hypothesis that the means under neutral and the corresponding replication (Able = 2 or Able = 8) mean are equal.

We next compare the beliefs provided by Allocators with those of neutral observers, to see whether $\hat{p}' \leq \hat{p} \leq \hat{p}^N$ holds in our new study. Table 3 compares the estimates of unbiased beliefs we obtained in our neutral treatment with the estimates provided by both *Able = 2* and *Able = 8* Allocators in our replication. The final two columns report statistical tests of the differences between neutral observers' beliefs and those of the two types of Allocators. The neutral and unbiased estimate is 44 percent for *Is Corrupt* and 46 percent for *%-Corrupt*.[27] Both of these are close to—and statistically indistinguishable from—the beliefs provided by Allocators with the high opportunity to take (Able = 8), again suggesting that these allocators exhibit no bias (i.e., $\hat{p} = \hat{p}^N$). Moreover, the estimates of 44 and 46 percent are very close to the actual frequency of corrupt choices by Sellers (42 percent). Thus, similarly to our first experiment, neutral observers provide fairly accurate estimates of behavior. In contrast, the mean beliefs provided by Allocators in the *Able = 2* treatment are considerably more positive than the beliefs provided by neutral observers and these differences are statistically significant.[28]

The results of this study provide support for our interpretation of a key distinction between relative versus absolute strategic cynicism in our study and in the one of Di Tella, et al., and for the theoretical model we used to account for these observations. Comparing the

[27] There is no significant difference between the beliefs of neutral subjects who received the instructions of an Able=2 Allocator and those who received instructions of an Able=8 Allocator.

[28] Non-parametric Wilcoxon rank-sum tests yield very similar results; p-values for Is Corrupt are 0.006 for the comparison between the neutral and the Able=2 Treatment and 0.574 for the comparison between the neutral and the Able=8 Treatment; p-values for %-Corrupt are 0.019 for the comparison between the neutral and the Able=2 Treatment and 0.835 for the comparison between the neutral and the Able=8 Treatment.

behavior of Allocators with a high and low taking opportunity, we observe that the former hold *relatively* more cynical beliefs—i.e., that $\hat{p}' \leq \hat{p}$, as in Proposition 2. We also observe that, in absolute terms, the estimates provided by Allocators do not exhibit a tendency toward cynicism, relative to the unbiased estimates of neutral observers—i.e., $\hat{p}' \leq \hat{p} \leq \hat{p}^N$, as in Propositions 1 and 3.

5. Conclusion

This paper studies whether the tendency to manipulate one's beliefs self-servingly extends to strategic cynicism, whereby an individual views her opponents' likely actions negatively when doing so can justify acting in a self-interested manner. We begin with a laboratory experiment that compares the beliefs of strategic players motivated to engage in strategic cynicism with the beliefs of neutral observers not incentivized to engage in any belief manipulation. We find no evidence that strategic actors manipulate their beliefs regarding opponents' behavior, thus seemingly contradicting the hypothesis of strategic cynicism, at least in *absolute* terms.

We then attempt to reconcile this observation with the results from Di Tella, et al. (2015), who find evidence of strategic belief manipulation, whereby subjects with greater opportunity to take money from another person are more cynical about the counterpart's likely behavior. This provides evidence of *relative* strategic cynicism, meaning that individuals become comparatively more cynical about their opponents when doing so justifies more self-interested behavior. However, Di Tella, et al.'s, data reveal very little evidence of strategic cynicism in absolute terms. Thus, one possible interpretation of the apparent discrepancy is that individuals exhibit relatively more pessimistic beliefs regarding the behavior of their counterparts when they stand to gain more from doing so, but that, in absolute terms, their beliefs will tend toward positivity rather than cynicism.

We show that this interpretation is consistent with a simple model—similar to the one that Di Tella, et al., use to motivate their study—in which individuals enjoy giving more if they believe that the beneficiaries are nicer. If we allow individuals to manipulate their beliefs with some cost for doing so, then the model can explain both of the above patterns. People benefit from thinking that they are acting toward kind others, but will believe these others to be less kind when greater cynicism lowers the costs of acting self-interestedly. Admittedly, this model is not general, but it provides a useful framework for reconciling the results of the two studies.

To investigate this interpretation, we conducted a novel experimental test. Specifically, we first replicated Di Tella, et al.'s, experiment and then extended it to obtain new measures of unbiased beliefs from neutral observers. Our replication confirms Di Tella, et al.'s, observation of relative strategic cynicism. Nevertheless, we also find that any absolute bias in beliefs seems to lie in the direction of too much positivity, rather than cynicism, about counterparts' behavior, especially by those with a limited opportunity to act self-interestedly.

We thus find no evidence—across many comparisons—that decision makers justify treating another person unfairly by self-servingly adopting the belief that the counterpart herself intends to act more egoistically than is actually the case. Instead, in terms of an absolute level of bias, we find evidence for another form of motivated belief; namely, in the kinds of interactions we study here, individuals seem motivated to convince themselves of the deservingness of the counterpart, and end up with beliefs that are often too positive. The finding that people are too positive about other players' kindness supports a general tendency for distorted beliefs to lie in the direction of positivity and optimism rather than the opposite. In fact, a broad view of the literature suggests that there is little evidence that people systematically bias their beliefs in a negative direction. While it is certainly unreasonable to rule out the possibility that there are contexts in which people may also engage in such cynical self-deception—and that, in absolute terms, this may even occur in strategic settings—our analysis suggests that such a tendency is limited, at least in some contexts.

References

Akerlof, G. A., and Dickens, W. T. (1982) "The Economic Consequences of Cognitive Dissonance," American Economic Review, 72(3): 307–319.

Akerlof, G. A., and Kranton, R. E. (2000) "Economics and Identity," Quarterly Journal of Economics 115(3), 715-753.

Andreoni, J., and Sanchez, A. (2014) "Do Beliefs Justify Actions or Do Actions Justify Beliefs? An Experiment on Stated Beliefs, Revealed Beliefs, and Social-Image Motivation," working paper.

Babcock, L., and Loewenstein, G. (1997) "Explaining Bargaining Impasse: The Role of Self-Serving Biases," Journal of Economic Perspectives 11(1): 109–126.

Bartling, B., and Özdemir, Y. (2017) "The Limits to Moral Erosion in Markets: Social Norms and the Replacement Excuse," working paper.

Bénabou, R. (2013) "Groupthink: Collective Delusions in Organizations and Markets," Review of Economic Studies, 80: 429-462.

Bénabou, R., and Tirole, J. (2006) "Incentives and Prosocial Behavior," American Economic Review, 96(5): 1652-1678.

Bénabou, R., and Tirole, J. (2011). "Identity, Morals, and Taboos: Beliefs as Assets," Quarterly Journal of Economics, 126: 805-855.

Bénabou, R., and Tirole, J. (2016) "Mindful Economics: The Production, Consumption, and Value of Beliefs," Journal of Economic Perspectives, 30(3): 141-164.

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. (2002) "Self-Signaling and Diagnostic Utility in Everyday Decision-Making," in Brocas, I. and Carrillo, J. eds., Collected Essays in Psychology and Economics, Oxford, UK: Oxford University Press.

Bodner, R., and Prelec, D. (2003) "Self-Signaling and Diagnostic Utility in Everyday Decision-Making," in Loewenstein, G., Read, D. and Baumeister, R.F. eds., Time and Decision, New York: Russell Sage Press.

Brunnermeier, M. K., and Parker, J. A. (2005) "Optimal Expectations," American Economic Review 95(4): 1092-1118.

Camerer, C., Loewenstein, G., and Weber, M. (1989) "The Curse of Knowledge in Economic Settings: An Experimental Analysis," Journal of Political Economy, 97(5): 1232-1254.

Chen, Z., and Gesche, T. (2017) "Persistent Bias in Advice-Giving," working paper.

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

Di Tella, R., Perez-Truglia, R., Babino, A., and Sigman, M. (2015) "Conveniently Upset: Avoiding Altruism by Distorting Beliefs about Others' Altruism," American Economic Review, 105(11): 3416-42.

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): 114-138.

Exley, C. (2016) "Excusing Selfishness in Charitable Giving: The Role of Risk," Review of Economics Studies, 83(2): 587-628.

Fischbacher, U. (2007) "z-Tree: Zurich toolbox for ready-made economic experiments," Experimental Economics, 10(2): 171–178.

Fischbacher, U., Gächter, S., and Fehr, E. (2001) "Are people conditionally cooperative? Evidence from a public goods experiment," Economics letters 71(3): 397-404.

Fischbacher, U., and Gächter, S. (2010) "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods experiments," American Economic Review, 100(1): 541-556.

Gino, F., Norton, M. I., and Weber, R. A. (2016) "Motivated Bayesians: Feeling Moral While Acting Egoistically," Journal of Economic Perspectives, 30(3): 189-212.

Gneezy, U., Saccardo, S., Serra-Garcia, M., and Van Veldhuizen, R. (2018) "Bribing the Self," working paper.

Grossman, Z. J., and van der Weele, J. (2017) "Self-Image and Willful Ignorance in Social Decisions," Journal of the European Economic Association, 15(1): 173–217.

Haisley, E., and Weber, R. A. (2010) "Self-serving interpretations of ambiguity in other-regarding behavior," Games and Economic Behavior, 68(2): 634-645.

Hamman, J., Loewenstein, G., and Weber, R. A. (2010) "Self-interest through delegation: An additional rationale for the principal-agent relationship," American Economic Review, 100(4): 1826-1846.

Irwin, F. W. (1953) "Stated Expectations as Functions of Probability and Desirability of Outcomes," Journal of Personality, 21(3): 329-335.

Jecker, J., and Landy, D. (1969) "Liking a person as a function of doing him a favor," Human Relations, 22: 371-378.

Konow, J. (2000) "Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions," American Economic Review, 90(4): 1072-1091.

Kunda, Z. (1990) "The Case for Motivated Reasoning," Psychological Bulletin, 108(3): 480–98.

Langer, E. J. (1975) "The Illusion of Control," Journal of Personality and Social Psychology, 32(2): 311-328.

Lerner, M. J. (1980) "The Belief in a Just World: A Fundamental Delusion." New York: Plenum Press.

Levine, D. K. (1998) "Modeling Altruism and Spitefulness in Experiments," Review of Economic Dynamics, 1: 593-622.

Madarász, K. (2012) "Information Projection: Model and Applications," Review of Economic Studies, 79: 961-985.

Mayraz, G. (2013) "Wishful Thinking," working paper.

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.

Möbius, M. M., Niederle, M., Niehaus, P., and Rosenblat, T. S. (2017) "Managing Self-Confidence," working paper.

OECD (2018) "Purchasing power parities (PPP) (indicator)," Accessed on 21 August 2018.

Quatrone, 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): 237-248.

Rabin, M. (1993) "Incorporating Fairness into Game Theory and Economics," American Economic Review, 83(5): 1281-1302.

Rabin, M. (1994) "Cognitive dissonance and social change," Journal of Economic Behavior and Organization, 23: 177-194.

Schopler, J., and Compere, J. S. (1971) "Effects of being kind or harsh to another on liking," Journal of Personality and Social Psychology, 20(2): 155-159.

Schwardmann, P., and van der Weele, J. (2016) "Deception and Self-Deception," working paper.

Shalvi, S., Gino, F., Barkan, R., and Ayal S. (2015) "Self-serving Justifications: Doing Wrong and Feeling Moral," Current Directions in Psychological Science 24(2): 125–130.

Svenson, O. (1981) "Are we all less risky and more skillful than our fellow drivers?," Acta Psychologica, 47: 143-148.

Taylor, S. E., and Brown, J. D. (1988) "Illusion and Well-Being: A Social Psychological Perspective on Mental Health," Psychological Bulletin, 103(2): 193-210.

Trivers, R. (2011) "The Folly of Fools: The Logic of Deceit and Self-Deception in Human Life," New York: Basic Books.

van der Weele, J. Kulisa, J., Kosfeld, M., and Friebel, G. (2014) "Resisting Moral Wiggle Room: How Robust Is Reciprocal Behavior?," American Economic Journal: Microeconomics, 6(3): 256-264.

Weinstein, N. D. (1980) "Unrealistic optimism about future life events," Journal of Personality and Social Psychology, 39(5): 806-820.

Weinstein, N. D. (1989) "Optimistic Biases About Personal Risks," Science, 246: 1232-1233.

Zimmerman, F. (2018) "The Dynamics of Motivated Beliefs," working paper.

Appendix A: Additional results

Figure A1 – Distribution of Player A choices in Experiment 1

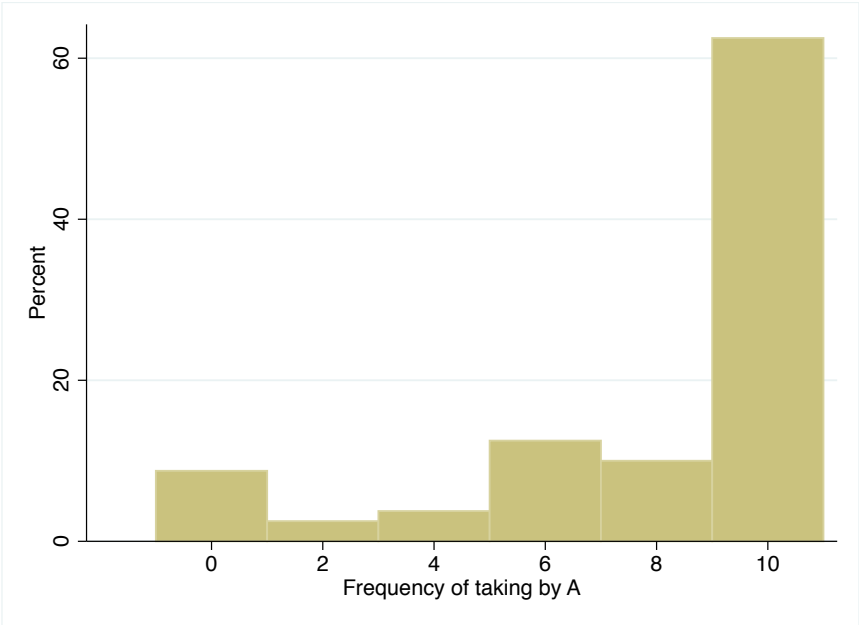


Figure A2 – Distribution of Player B choices in Experiment 2

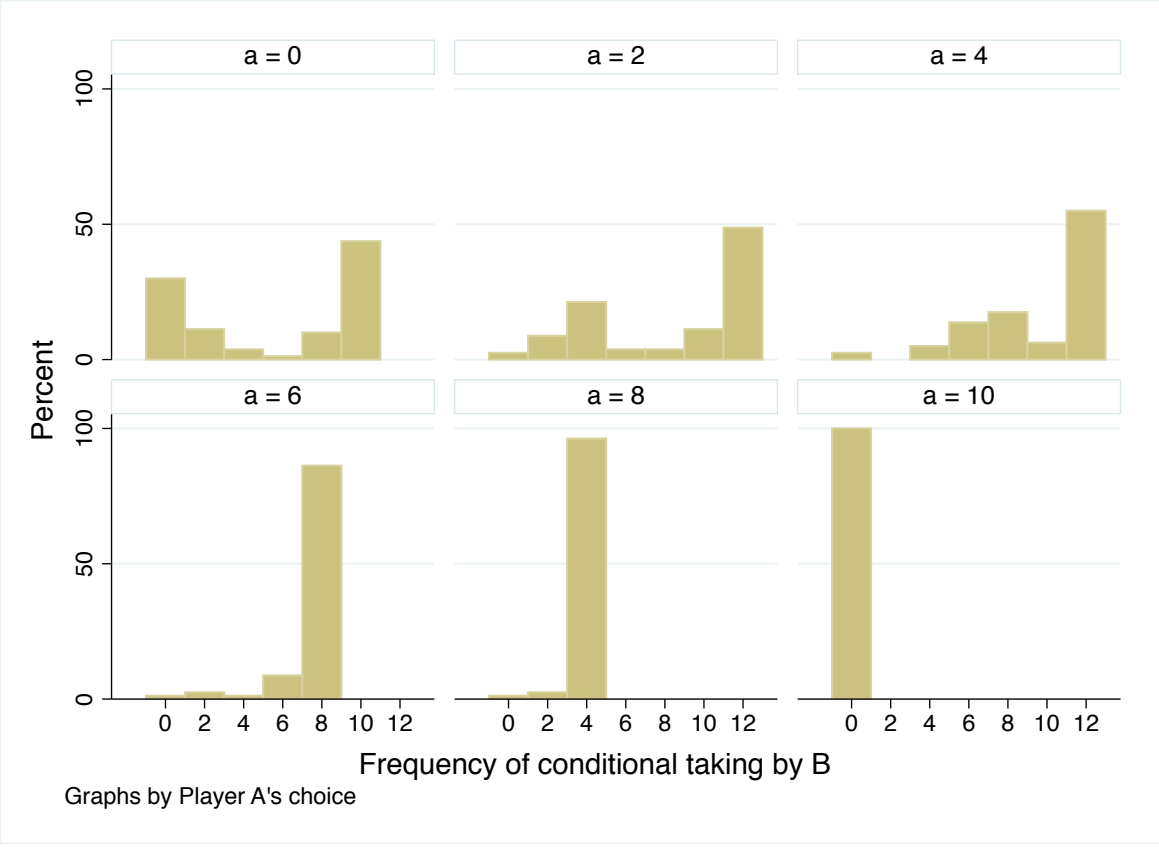
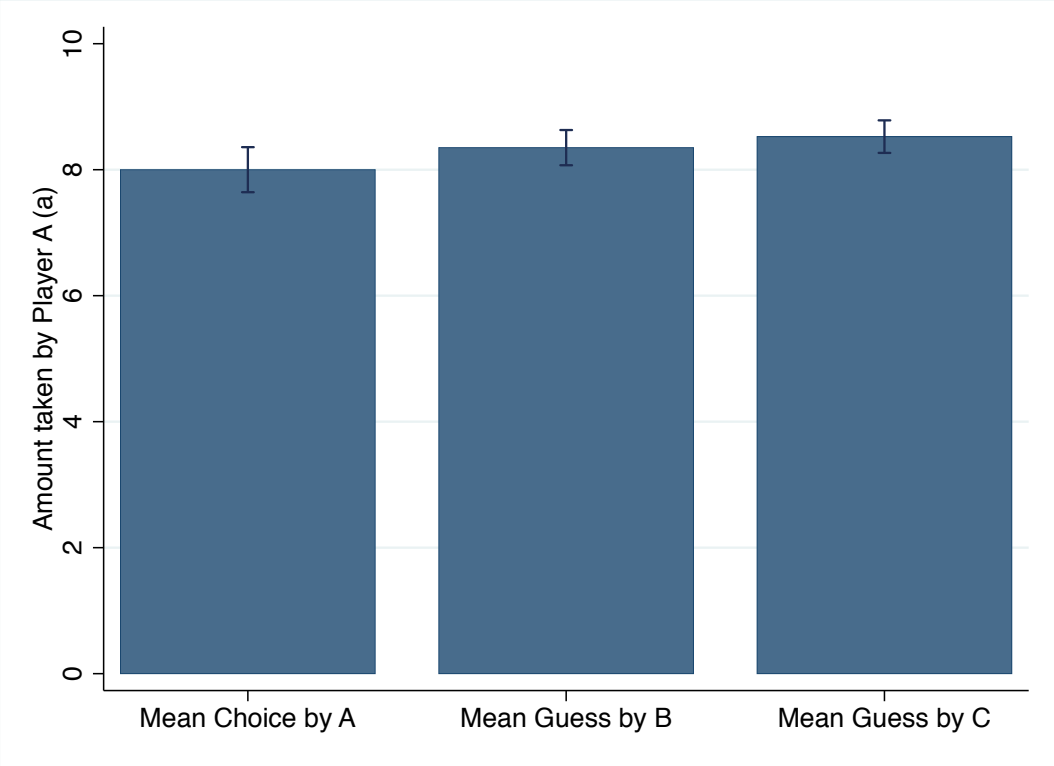


Figure A3 – Taking by Player A and corresponding beliefs of Players B and C



Appendix B: Proofs

Proposition 1: For any $(x, \hat{p})$ *in* $argmax_{x\in[0,K],\hat{p}\in[0,1]}\, \mathrm{U}(x, \hat{p})$, $\hat{p} \leq p$.

Proof: For any $\hat{p} \in (p, 1]$ and any $x \in [0, K]$ we have that $U(x, p) > U(x, \hat{p})$:

$$U(x,p) - U(x,\hat{p}) = \left(E_p(d) - E_{\hat{p}}(d)\right) v(1-x) + \left(C(\hat{p}-p) - C(0)\right) + (P(0) - P(\hat{p}-p))$$

First, note that $\left(E_p(d) - E_{\hat{p}}(d)\right) v(1-x) > 0$ as $E_p(d) > E_{\hat{p}}(d)$ and $v(.) > 0$. Second, we have $C(\hat{p}-p) - C(0) > 0$ as $C'(.) > 0$. Third, $P'(.) < 0$ implies that $P(0) - P(\hat{p}-p) > 0$. Therefore $U(x,p) - U(x,\hat{p}) > 0$, and $\hat{p} \in (p, 1]$ cannot be a solution to (1).

Proposition 2: Take K, K' *in* $(0,1]$ *with* $K' < K$ *and suppose that there is a unique solution to (1) for both* K *and* K'*, then* $\hat{p}' \leq \hat{p}$.

Proof: Define $(x, \hat{p}) = argmax_{x\in[0,K],\hat{p}\in[0,1]}\, \mathrm{U}(x, \hat{p})$ and $(x', \hat{p}') = argmax_{x\in[0,K\prime],\hat{p}\in[0,1]}\, \mathrm{U}(x, \hat{p})$.

Case i) Suppose $x \leq K'$, then $(x, \hat{p})$ is the solution of $max_{x\in[0,K\prime],\hat{p}\in[0,1]}\, U(x, \hat{p})$, so $\hat{p}' = \hat{p}$.

Case ii) Suppose $x(K) > K'$. Then $x(K') < x(K)$ (and $1 - x(K') > 1 - x(K)$). Note that the relevant Karush–Kuhn–Tucker conditions for the problem are (Proposition 1 implies that the condition $\hat{p} \leq 1$ is not binding):

$$-\mathrm{P}'(p-\hat{p}) - (d_H - d_L)v(1-x) + C'(p-\hat{p}) + \lambda = 0 \text{ (I)}$$
$$\lambda\hat{p} = 0 \text{ (II)}$$
$$\hat{p} \geq 0 \text{ (III)}$$
$$\lambda \geq 0 \text{ (IV)}$$

Case iia) Suppose $\hat{p}' = 0$, then $\hat{p}' \leq \hat{p}$ due to condition (III).

Case iib) Suppose $\hat{p} = 0$. Then (I) implies

$$(d_H - d_L)v(1-x) \geq C'(p) - \mathrm{P}'(p) \text{ (I')}$$

Note that $(d_H - d_L)v(1-x') > (d_H - d_L)v(1-x)$ as $v'(.) > 0$. This, together with (I') and the assumptions on C and P imply $(d_H - d_L)v(1-x') > C'(p) - \mathrm{P}'(p) \geq C'(p-\hat{p}') - \mathrm{P}'(p-\hat{p}')$. Then by (I), $\lambda' > 0$. Then by (II) $\hat{p}' = 0$.

Case iic) Suppose $\hat{p}, \hat{p}' > 0$. Then by (II) $\lambda = \lambda' = 0$. Then (I) simplifies to:

$$(d_H - d_L)v(1-x) = C'(p-\hat{p}) - \mathrm{P}'(p-\hat{p}) \text{ (I'')}$$
$$(d_H - d_L)v(1-x') = C'(p-\hat{p}') - \mathrm{P}'(p-\hat{p}') \text{ (I''')}$$

Combining (I''), (I''') and $(d_H - d_L)v(1-x') > (d_H - d_L)v(1-x)$ implies:

$$C'(p-\hat{p}') - \mathrm{P}'(p-\hat{p}') > C'(p-\hat{p}') - \mathrm{P}'(p-\hat{p}) \text{ } (V)$$

(V) together with $P'(.) < 0$, $P''(.) \leq 0$, $C'(.) > 0$ and $C''(.) > 0$ implies $\hat{p}' \leq \hat{p}$.

Proposition 3: $\hat{p}^N = p$ *is the unique solution to* $\max_{\hat{p}^N\in[0,1]} U(\hat{p}^N)$.

Proof: For any $\hat{p}^N \in (p, 1]$ we have that $U(p) - U(\hat{p}^N) = P(0) - P(\hat{p}^N - p) + C(\hat{p}^N - p) - C(0) > 0$ due to $C'(.) > 0$ and $P'(.) < 0$.

For any $\hat{p}^N \in [0, p)$ we have that $U(p) - U(\hat{p}^N) = P(0) - P(p - \hat{p}^N) + C(p - \hat{p}^N) - C(0) > 0$ due to $C'(.) > 0$ and $P'(.) < 0$.