Heterogeneity in lies and lying preferences

Katharina A. Janezic*

December 31, 2020

Job Market Paper

[Click here for most recent version]

Abstract

How can we take heterogeneity of both lies and lying preferences into account when analysing and predicting individual lying decisions? This paper provides a unifying framework to analyse and predict lying behaviour. At the theoretical level, the framework provides a complete categorisation of lie types and a simple, testable model. At the experimental level, the framework guides a novel experimental design which allows the observation of individual choices, can assess variation of behaviour across lie types as well as disentangle the effects of lie types from standard social preferences. I find that varying the lie type has a large and significant impact on behaviour and that this behaviour is highly heterogeneous across individuals. In order to quantify the heterogeneity of decision-makers, I employ an unsupervised machine learning algorithm to identify coherent types of decision-makers and their distribution in the data. I show that the identified types are meaningful as accounting for them in an out-of-sample forecasting exercise significantly improves the precision of the forecast. In contrast, I find that social preferences have only limited explanatory power for lying preferences. Finally, to show that my framework can be used for the purpose of model building and prediction, I present a parametric version of the framework and calibrate it at the individual and the group level.

JEL Codes: C91; D91.
Keywords: honesty - lying - structural estimation - heterogeneity - machine learning - private information - experiment - social preferences.

*Universitat Pompeu Fabra and Barcelona GSE. E-mail: katharina.janezic@upf.edu. Website: https://www.katharinajanezic.com

Acknowledgements. I would like to thank Larbi Alaoui and Jose Apesteguia for their invaluable comments, support and advice. I would also like to thank Antonio Penta for his support and helpful comments. Further, I thank Johannes Abeler, Pierpaolo Battigalli, Antonio Cabrales, Colin Camerer, Francesco Cerigioni, Andrew Ellis, Lukas Hoesch, Navin Kartik, Gaël Le Mens, Rosemarie Nagel, Yusufcan Masatlioglu, Pedro Rey Biel and Balazs Szentes as well as seminar participants at Universitat Pompeu Fabra and the London School of Economics for their useful feedback and comments.

1. Introduction

For centuries, philosophers and lawmakers have concerned themselves with the concept of lying, debating questions such as whether people should lie, which lies ought to be punishable and how they can be prevented. More recently, economists have joined the debate (see for example Gneezy (2005), Abeler, Nosenzo, and Raymond (2019)). Despite this continued interest, we still cannot anticipate when someone will lie, often because individuals' preferences, as well as the lies themselves, differ substantially from each other. How can we take heterogeneity of both lies and lying preferences into account when analysing and predicting individual lying decisions?

In this paper, I provide a unifying framework to analyse individual lying preferences. Previous papers have shown that people are highly heterogeneous when it comes to their lying behaviour (see for example Gibson, Tanner, and Wagner (2013)). At the same time, most existing research has used methods that analyse lying decisions at the aggregate level or has elicited individual preferences in environments where they could not be separated from belief effects. Due to the existence of heterogeneity, analysing lying decisions at the aggregate level might come at the cost of individual differences averaging each other out and preventing us from understanding lying behaviour. Similarly, if beliefs about a possible response to (lying) decisions and lying preferences are conflated, it is challenging to elicit lying preferences from observed behaviour. For this reason, it is crucial to research lying preferences at the individual level without confounding beliefs. Moreover, not only the decision-makers are heterogeneous, but so are the lies themselves. Lies differ with respect to their consequences and can be classified into types accordingly. My framework provides a toolkit to help us understand how these heterogeneities interact. The theoretical framework provides a classification of lies as well as a simple model with testable predictions for individual behaviour. At the experimental level, I introduce a novel experimental design that makes it possible to observe lies at the individual level while removing confounding factors such as first or higher order beliefs. In addition to controlling for such beliefs, my experimental design allows the separation of lying preferences from social preferences. Finally, to show that the framework can be used for the purpose of model building and prediction, I present a parametric version of the framework and calibrate it at the individual and the group level.

My unifying framework contributes to the literature along four dimensions. First, at the theoretical level, the framework provides a classification of the lie type space and a simple non-parametric model. Formally, lie types in this paper are defined in terms of lies' consequences on the monetary payoff of a decision-maker and another person who is affected by the lie, which follows the approach first introduced by Gneezy (2005). A

complete characterisation is a natural prerequisite for modelling heterogeneity of lies in a systematic manner. The model provides guidance on how to examine behaviour in the context of these lies. It is a private information model with the following set-up: A decision-maker privately observes the true state of the world and is asked to report it. She can report any state from a set of potential states. The decision-maker and another person, the 'partner', are affected by the reported state in that they receive monetary payoffs connected to that state. There are no strategic concerns, as the partner cannot influence the outcome and the report is unverifiable. Preferences are heterogeneous with respect to the importance that the decision-makers attach to their own and the partner's payoffs as well as with respect to the cost of lying. Behaviour is characterised by a single crossing property: if a subject preferred to lie, she should not revert back to telling the truth when either of the payoffs associated with lying increases relative to the observed lie. This set-up allows the clear definition and systematic variation of lie types needed to study and model their impact on lying behaviour and provides testable predictions that help guide the experimental design. Due to the framework's properties, 'regions of inference' can be constructed in which the decision-maker is expected to lie. These regions provide a non-parametric tool to predict behaviour and can be used to falsify the model.

Second, the paper contributes along the experimental dimension by introducing a novel experimental design that allows the researcher to observe individual choices without suffering from the confounding factors of first or higher order beliefs. It thereby combines the advantages from the two major experimental designs used in the lying literature, die rolling and sender-receiver games, while avoiding their drawbacks. Strategic beliefs make it difficult to elicit lying preferences from observed decisions for the following reason: if a decision-maker decides between telling the truth or a lie to another person and this other person does not know whether it was the truth but can react to it, then expectations of the other person's response will feed back into the decision-maker's decision of whether to lie or not. For example, if she thinks that the other person will not trust her even if she tells the truth, she might tell the truth not because of lying preferences but because of strategic considerations. Eliciting lying preferences from observed behaviour is then very challenging. That beliefs are inconsequential in my setting in conjunction with knowledge of individual choices makes it possible to pinpoint the effect of lie types on behaviour. The behaviour elicited in the experiment is used to examine the descriptive and predictive power of the framework. The experimental results show that varying the type of lie has a large and significant impact on subjects' behaviour. Moreover, this effect goes beyond differences in payoffs which suggests that the lie type plays a psychological role in itself. Importantly, accounting for lie types improves the performance of prediction exercises, as measured by a reduction in the mean squared forecast error (MSE), by more than 21.5 percent.

The results demonstrate the importance of accounting for heterogeneity of lie types in models and in experiments. I show that the impact of lie types is highly heterogeneous across subjects and several behavioural types can be identified in the data. Specifically, I employ a machine learning approach which combines a principal component analysis with a k-means analysis in order to group subjects based on their decisions throughout the experiment. The k-means clustering analysis is a statistical pattern recognition tool that, in economics, is commonly used in time-series econometrics (see for example Falat and Pancikova (2015), Bagnall et al. (2003), Focardi and Intertek Group (2001-2004)). The algorithm identifies six separate behavioural groups which differ with respect to the number and type of lies that they tell in the experiment. Not only can the method uncover heterogeneity but also quantify it. I show that the types of behaviour identified by the algorithm are systematic and meaningful in the sense that subjects within groups make similar choices to each other and that a narrative explaining those choices can be found. Exploiting the uncovered heterogeneity improves out-of-sample forecasting accuracy by more than 60 percent.

Third, the experimental design makes it possible to separate lying preferences from the potential confound of standard social preferences, such as altruism or inequality aversion. In principle, differences in behaviour across lie types could be fully explained by social preferences as lie types are categorised based on variations in monetary outcomes. If that were the case, we should expect that social preferences can predict how subjects behave in the lying game and a novel lying framework would not be necessary as we could simply rely on the social preference literature to explain lying behaviour in the context of lie types. In order to directly contrast lying preferences with social preferences, the experiment contains two separate but nearly identical games. In the first, called the lying game, a large number of lie-truth choices per individual is elicited. In the second game, subjects choose between the same alternatives as in the lying game but alternatives are no longer classified as truths or lies as there is no truth benchmark. My contribution here is the direct comparison between lying preferences and social preferences. This is possible due to choices being observable at the individual level in the lying game and in the corresponding social preference game. I find that knowing how subjects respond in a social preference game does not help to predict behaviour in the lie setting despite both choice settings being identical to each other except for the choice objects' labels (unlabelled versus truth and lie labels). Thus, the finding confirms that a lying framework is indeed necessary to describe lying preferences even in situations where lie types are fully captured by payoff differences affecting the decision-maker and another person. This further confirms that the effect of lie types goes beyond that of payoff consequences and that they have an additional psychological impact on the decision-maker.

Fourth, having shown that, first, lie types matter and second, that people respond systematically to these lie types, I provide a parametric model that can capture these lying preferences. Specifically, I propose a simple parametric version of the general model that performs well for the data generated by the experiment. Based on the behaviour elicited in the experiment, I calibrate the model's parameters using a maximum likelihood estimation (MLE) approach. I find substantial variation in the estimated lie regions across subjects, in line with the hypothesised decision-maker types.

The theoretical and experimental results highlight that a unifying framework which can account for heterogeneity of both lies and decision-maker's preferences vastly improves our ability to capture, understand and, most importantly, predict lying behaviour.

The rest of the paper is structured as follows: The next section discusses the related literature. Section 2 presents a model and several properties for lying behaviour under lie types. Section 3 introduces the design and the logistics of the experiment. The experimental results are presented in Sections 4 and 5. Section 6 introduces a parametrised utility function in line with the framework and discusses the results of a calibration exercise. Section 7 concludes the paper.

1.1. Literature Review

In his seminal paper, Gneezy (2005) defines lies based on their monetary consequences and Erat and Gneezy (2012) extend this to a more formal definition of lie types. Specifically, each lie type is defined by two outcomes, the monetary payoff of the decision-maker and that of the partner, relative to the payoffs associated with telling the truth. Since then, the literature has examined many different aspects of lying such as the existence of moral costs of lying (Kartik (2009), Gibson, Tanner, and Wagner (2013)), measures of lying (Fischbacher and Föllmi-Heusi (2013), Gneezy, Rockenbach, and Serra-Garcia (2013)), or the role of deliberation time (Capraro (2017), Lohse, Simon, and Konrad (2018)). While these studies have focused on different explanations of lying preferences, they have in common that they focus their attention on egoistic lies i.e. those lies that benefit the liar at the expense of someone else.

A smaller literature also examines other types of lies such as self-serving (only affect the liar) or mutually beneficial lies (beneficial to all people affected by the lie), and a handful have considered multiple types of lies simultaneously (Erat and Gneezy (2012), Levine and Schweitzer (2014), Biziou-van Pol, Haenen, Novaro, Occhipinti Liberman, and Capraro (2015)). Erat and Gneezy (2012) establish that behaviour varies across the lie types studied. Biziou-van Pol, Haenen, Novaro, Occhipinti Liberman, and Capraro (2015)

study both altruistic lies (lies where the decision-maker incurs a loss and the partner a gain relative to the truth) and mutually beneficial lies. A puzzle in this literature has been the large differences in the percentages of liars across papers. Even in papers that compare lies of the same lie types, this puzzle persists (compare for example Biziou-van Pol, Haenen, Novaro, Occhipinti Liberman, and Capraro (2015), Erat and Gneezy (2012) and Hurkens and Kartik (2009)). I contribute to this literature by utilising a large set of lies in the experimental set-up. While most papers use one to five lie specifications when examining lying behaviour, this paper uses 60. I can thereby systematically assess how behaviour varies across the complete space of lies and show that the type but also the specifics of each lie matter in determining how many people will tell that lie. This underlines the need for systematic analyses as compared to standard practices where one lie is taken as representative of the whole space of lies.

Two experimental paradigms are particularly prominent in the analysis of lying preferences: the die rolling paradigm (Fischbacher and Föllmi-Heusi (2013)) and sender-receiver games (Gneezy (2005), Hurkens and Kartik (2009)). In the die rolling experiments, subjects are asked to privately roll a die and to report the outcome. Monetary payoffs are increasing in the outcome of the die so that subjects have an incentive to report higher numbers irrespective of the true outcome. They thus have an incentive to lie. The empirical distribution of reported outcomes can then be statistically compared to the theoretical distribution which, under the assumption that no one lies, predicts that in sufficiently large samples, all six numbers come up with equal probability. If there exists a statistically significant difference between the theoretical and the empirical distribution, this can be attributed to lying. A popular variation of this methodology is the coin flipping paradigm (see for example Abeler, Becker, and Falk (2014)), where a coin is flipped instead. The die rolling paradigm captures lying preferences at the group level and is belief free in the sense that there are no reputation concerns and no one other than the decision-maker is directly affected by the lie. On the other hand, sender-receiver games can measure individual preferences but introduce confounds via first or higher order beliefs. Here, a sender decides whether to send a truthful or a dishonest message to the receiver. The receiver then responds with an action, based on beliefs of whether the message was truthful or not. Senders anticipate the response and thus include beliefs about the possible responses in their initial decision. My experiment contributes to the literature by being able to combine the advantages of both of these experimental designs. It captures individual preferences without suffering from the confound of belief effects.

Kerschbamer, Neururer, and Gruber (2019) as well as Hurkens and Kartik (2009) and Biziou-van Pol, Haenen, Novaro, Occhipinti Liberman, and Capraro (2015) study the connection between social and lying preferences. Most papers that examine this potential

connection elicit social preferences in a setting different from the lying experiment that they employ. For example, Kerschbamer, Neururer, and Gruber (2019) use the *Equality Equivalence Test* and Biziou-van Pol et al. (2015) use both a dictator and a prisoner's dilemma game in order to elicit social preferences. Their analysis is at the group level. Hurkens and Kartik (2009) employ a sender-receiver game and ask all senders to participate in a dictator game. While they find no statistically significant difference, the authors acknowledge that only 55% of senders believed that the receivers would believe their message so that it is difficult to disentangle trust, strategic and distributional preferences from lying preferences. I contribute here by being able to abstract from such confounding factors and by directly contrasting each individual's lying choices with distributional choices rather than relying on group averages.

Abeler, Nosenzo, and Raymond (2019) present an extensive meta study on lying. They provide insights into which concerns enter the decision to lie. As a large number of past papers has employed the die rolling and coin flipping paradigms, the authors measure lying preferences at the group level. I complement their analysis by focusing on the individual level, instead. In addition, I contribute by examining lies other than self-serving lies (those lies that benefit the decision-maker and have no direct effect on someone else) and provide a machine learning approach to assess the heterogeneity of lying preferences.

2. Theoretical Framework

I consider a setting with two individuals, where one individual has to choose between telling the truth or lying. The first individual is called the *decision-maker* and the second is called the *partner*. The decision-maker reports a privately observed state; the report can either be truthful or a lie and is non-verifiable. The report has payoff consequences for both individuals. The partner is passive in the sense that he cannot influence the outcome but is affected by the decision that the first individual makes via the payoffs. As this paper focuses on intrinsic motives to lie, there is no strategic interaction between decision-makers and their partners. Therefore, beliefs over the partner's actions are inconsequential for the decision-making process. Furthermore, payoffs will be paid out with certainty and can therefore enter the decision-making process directly rather than in expectation.

I now present a framework and some basic properties that formalise behaviour under such a setting. These properties help explain lie behaviour and inform both the design of the experiment and the analysis of the experiment's results.

A decision-maker (DM) faces a decision problem $d \in \{1, \ldots, D\}$. For each decision problem d, there exists a finite set of potential states of the world $S_d = \{s_d^0, s_d^1, \ldots, s_d^M\}, M \geq 1$, with typical element s_d^m, $m = 0, 1, \ldots, M$. Each decision problem d contains one *true*

state, which is denoted by s_d^0. All other states in S_d are *untrue states*. The DM privately observes S_d with true state s_d^0. Any state in S_d can be publicly reported as the true state by the DM, irrespective of whether it is the true state s_d^0 or one of the untrue states. The states are payoff relevant in that any reported state s_d^m is tied to monetary payoffs $(x_d^m, y_d^m) \in \mathbb{R}_+^2$. The DM's payoff is given by x_d^m and the partner's payoff is y_d^m. The DM knows the mapping between states and payoffs. The DM chooses to report the state that maximises his utility, which depends on both the true and the reported state. The state that is reported is denoted by r_d. We say that a DM decides to lie whenever $r_d \neq s_d^0$.

Table 1 defines the possible lie types when s_d^0 is the true state and s_d^m is any other potential state. Lie types are defined by comparing the decision-maker's and the partner's payoffs from lying relative to telling the truth. To illustrate, imagine that a DM faces a decision problem d where the true state s_d^0 is associated with the payoff bundle $(x_d^0, y_d^0) = (5, 5)$. S_d contains one other, untrue state, s_d^1 with payoffs $(x_d^1, y_d^1) = (10, 10)$. If the DM reports s_d^0, he tells the truth and will obtain a payoff of $(5, 5)$ for himself and his partner. If, instead, he reports s_d^1, he tells a lie and obtains payoffs $(10, 10)$. Such a lie is called a *mutually beneficial lie (MBL)* as both affected individuals obtain a monetary benefit from the lie relative to the truth. Now imagine that S_d contains a further, untrue, state s_d^2 with payoffs $(x_d^2, y_d^2) = (15, 0)$. If the DM were to decide to report s_d^2, he would tell an *egoistic lie (EL)* as telling the lie increases his payoff but decreases the partner's payoff relative to the truth. If S_d also contains untrue state s_d^3 with payoffs $(x_d^3, y_d^3) = (0, 15)$ and the DM decides to report s_d^3, then he would tell an *altruistic lie (AL)* as the DM's payoff decreases but the partner's payoff increases relative to reporting the truth. Following this logic, there are nine lie types, as defined in Table 1.

Table 1: Lie Types

	$y_d^m > y_d^0$	$y_d^m = y_d^0$	$y_d^m < y_d^0$
$x_d^m > x_d^0$	Mutually beneficial (MBL)	Self-serving (SSL)	Egoistic (EL)
$x_d^m = x_d^0$	Weakly Altruistic (WAL)	Neutral (NL)	Harmful (HL)
$x_d^m < x_d^0$	Altruistic (AL)	Self-harming (SHL)	Mutually harmful (MHL)

Lie types are defined based on the payoff consequences of the lie relative to the truth for the decision-maker and the partner. A report is classified as a lie whenever $r_d \neq s_d^0$. x_d^0 stands for the decision-maker's payoff from reporting the true state s_d^0, x_d^m for the decision-maker's payoff from reporting an untrue state s_d^m. y_d^0 stands for the partner's payoff when the true state s_d^0 is reported and y_d^m for the partner's payoff when an untrue state s_d^m is reported.

I now define the DM's preferences. I allow for heterogeneity of lying preferences and represent this heterogeneity by a vector θ with N entries $\theta_n \in \mathbb{R}$. The elements of θ,

$\theta_1, \ldots, \theta_N$, refer to the different dimensions that enter the decision-making process such as the mental cost of lying, the strength of other-regarding preferences, or the effect of lie types. The vector θ thus defines those dimensions of the decision-making process that result in differences in lying preferences. The DM's preferences are described by the following utility function, with $m = 0, \ldots, M$:

$$u(s_d^m; s_d^0, \theta) := v(x_d^m, y_d^m, LT(x_d^m - x_d^0, y_d^m - y_d^0); \theta) - \mathbb{1}_{s_d^m \neq s_d^0} c_\theta$$

Component c_θ captures the psychological cost of lying, assumed to be constant for an individual but varying across individuals. Function $v(x_d^m, y_d^m, LT(x_d^m - x_d^0, y_d^m - y_d^0); \theta)$ captures several aspects of lying: the utility from payoff consequences for both the DM and the partner but also the effect of a lie type on utility. This last aspect is captured by $LT(x_d^m - x_d^0, y_d^m - y_d^0)$. Notice that both $LT(\cdot)$ and $v(\cdot)$ depend entirely on the monetary payoffs of a reported state and on the typology of lies, which is also consequence based. $LT(\cdot)$ can have both a positive or a negative effect on utility so that this can capture both utility as well as disutility across lie types. Importantly, utility itself is consequence based. The state that is publicly reported by the DM, r_d, is the state which maximises the DM's utility, taking into consideration whether reporting constitutes the truth or one of the nine lie types:

$$r_d \in \arg\max_{s_d^m \in S_d} u(s_d^m; s_d^0, \theta)$$

If S_d has only two elements, the choice is binary between telling a lie and telling the truth. If there are more than two states, the DM can decide between telling the truth (reporting s_d^0) or between telling one of several possible lies.

The framework's purpose is to guide the experimental design and the interpretation of the experimental results. As such, some identifying assumptions are needed in order to interpret observed behaviour effectively. The first identifying assumption (or property) is a tie breaking rule. It states that if the utilities from reporting the true state s_d^0 is larger or equal to the utilities of reporting untrue states s_d^m, then the true state is reported. Notice that when the utilities are equal to each other, the tie-breaking property guarantees that the true state will be reported (the case when they are unequal is already governed by the the fact that r_d is the arg max).

Property 1 Tie-breaking rule
For all $d \in \{1, \ldots, D\}$, if, for all $s_d^m \neq s_d^0$, $u(s_d^0; s_d^0, \theta) \geq u(s_d^m; s_d^0, \theta)$, then $r_d = s_d^0$.

The second identifying property states that the truth is weakly preferred to lying when there are no monetary consequences from lying relative to telling the truth.

PROPERTY 2 TRUTH IS WEAKLY PREFERRED
For all $d \in \{1, \ldots, D\}$ such that $x_d^0 = x_d^m$ and $y_d^0 = y_d^m, m \neq 0$, it holds that $u(s_d^0; s_d^0, \theta) \geq u(s_d^m; s_d^0, \theta)$.

For Property 2, when $u(s_d^0; s_d^0, \theta) = u(s_d^m; s_d^0, \theta)$ and the utilities are larger than those of any other state, Property 1 ensures that the true state is reported. Note that Property 1 governs situations with equality of utilities while Property 2 governs situations with equality of monetary payoffs. Together, these properties ensure that when a DM in the experiment reports an untrue state, reporting the untrue state must have generated strictly greater utility than reporting the truth. Observing an untrue state being reported thus reveals a strict preference for telling a lie.

Property 3 defines the heterogeneity of lying preferences in this framework. To ease notation, the difference in payoffs from lying and telling the truth will be given by: $\Delta x_d^m = x_d^m - x_d^0$ and $\Delta y_d^m = y_d^m - y_d^0$. They will be referred to as "relative payoffs of lying".

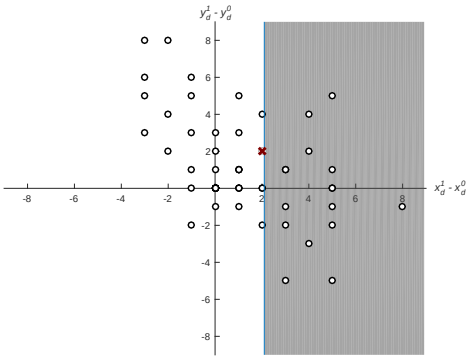
PROPERTY 3 REGIONS OF INFERENCE
For all decision problems $d, \bar{d} \in \{1, \ldots, D\}$ such that $r_d = s_d^m$ with states $s_d^m \neq s_d^0$ and $s_{\bar{d}}^{\bar{m}} \neq s_{\bar{d}}^0$, at least one of the following has to hold:
(1) if $\Delta x_{\bar{d}}^{\bar{m}} \geq \Delta x_d^m$ then $u(s_{\bar{d}}^{\bar{m}}; s_{\bar{d}}^0, \theta) > u(s_{\bar{d}}^0; s_{\bar{d}}^0, \theta)$.
(2) if $\Delta y_{\bar{d}}^{\bar{m}} \geq \Delta y_d^m$ then $u(s_{\bar{d}}^{\bar{m}}; s_{\bar{d}}^0, \theta) > u(s_{\bar{d}}^0; s_{\bar{d}}^0, \theta)$.
(3) There exists $\alpha_\theta \in (0, 1)$ such that if $\alpha_\theta \Delta x_{\bar{d}}^{\bar{m}} + (1 - \alpha_\theta)\Delta y_{\bar{d}}^{\bar{m}} \geq \alpha_\theta \Delta x_d^m + (1 - \alpha_\theta)\Delta y_d^m$ then $u(s_{\bar{d}}^{\bar{m}}; s_{\bar{d}}^0, \theta) > u(s_{\bar{d}}^0; s_{\bar{d}}^0, \theta)$.

Property 3 states that if a DM preferred to lie for a particular decision-problem, he will also lie for decision-problems in which the relative payoff(s) of lying are weakly larger relative to those of the decision-problem where the DM already lied. Thus, increasing the gains of lying guarantees that there is no reversal in his preferences once the DM starts to lie. This implies that there exists a well defined "lie region" and that once a DM has moved to the "lie region", they will remain in that region.

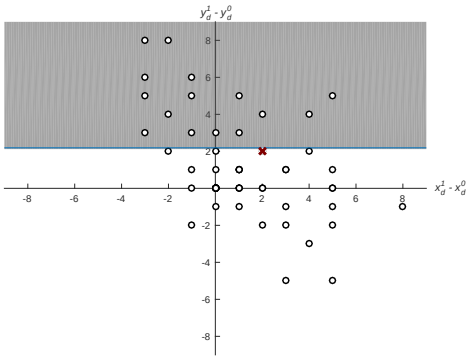
Depending on which of the sub-properties hold, "the gain" can refer to the DM's, the partner's or both relative payoffs. Property 3(1) implicitly assumes that the DM does not care about the partner's payoff as any increase in his own payoff realtive to the truth and

Figure 1: Visualisation of Property 3

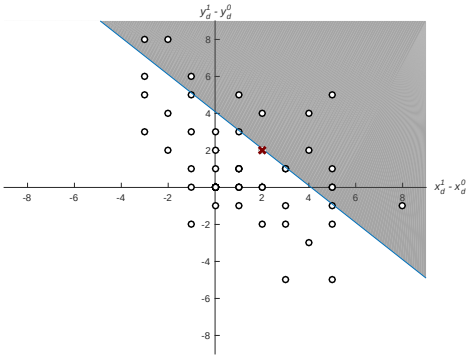
(a) Property 3(1)



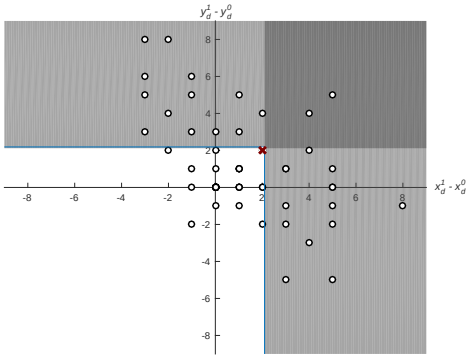
(b) Property 3(2)



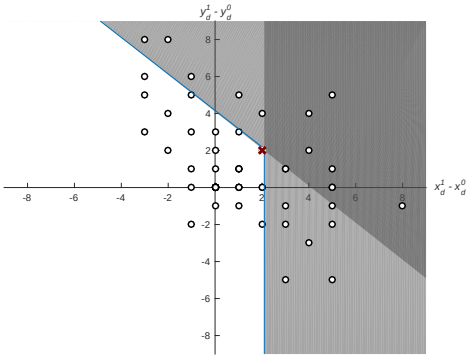
(c) Property 3(3), $\alpha_\theta = 0.5$



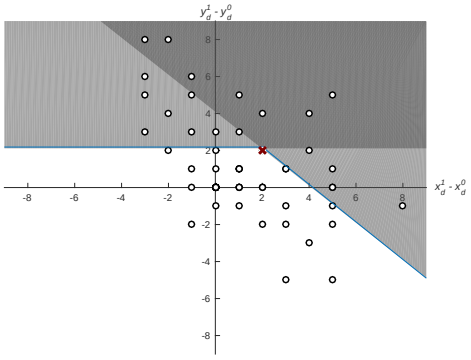
(d) Properties 3(1) and (2)



(e) Properties 3(1) and (3), $\alpha_\theta = 0.5$



(f) Properties 3(2) and (3), $\alpha_\theta = 0.5$



Implications of Property 3 visualised using an example. The red cross visualises the assumption that the DM lied for the payoff combination of $\Delta x_d = 2$, $\Delta y_d = 2$ where $\Delta x_d = x_d^1 - x_d^0$ and $\Delta y_d = y_d^1 - y_d^0$ are displayed on the horizontal and the vertical axis respectively. The grey areas indicate payoff combinations for which the DM has to lie, given Property 3. Darker grey tones for figures that display combinations of sub-properties indicate that sub-properties coincide with their prediction for this area.

the lie that he already told will lead him to continue to lie, irrespective of the change in the partner's payoff. Property 3(2) makes the opposite assumption in that the decision to lie is invariant to his own payoff but he will lie whenever the partner's payoff relative to the truth increases relative to the lie that he already told. Property 3(3) says that both changes in both payoffs matter. If the weighted sum of the payoffs relative to the truth payoffs is larger than the weighted sum of the relative payoffs of the lie that he already told, he will continue to lie. The parameter α_θ ensures that the payoffs can be weighted differently across individuals such that some DMs are allowed to care more about their own payoff than about the partner's payoff or vice versa. The reason why α_θ cannot take the extreme values of 0 and 1 is the following: if it were to take one of these values, it would subsume sub-properties (2) and (1) respectively. In such a case, we could simply use only sub-property (3) to describe all types. However, if that were the case, we could no longer have combinations of the sub-properties (see paragraph below). Thus, we need all three sub-properties to be separate from each other and therefore do not allow α_θ to be 0 or 1.

It is important to emphasise that while Property 3 only requires that one of the sub-properties (1) - (3) has to hold, the sub-properties can also hold simultaneously. In such a case, the implicit assumptions are valid only for parts of the payoff space. To illustrate, imagine that Property 3(1) and 3(3) hold simultaneously. Then the DM should lie whenever $\Delta x_{\bar{d}} > \Delta x_d$ and whenever $\alpha_\theta \Delta x_{\bar{d}}^{\bar{m}} + (1-\alpha_\theta)\Delta y_{\bar{d}}^{\bar{m}} > \alpha_\theta \Delta x_d^m + (1-\alpha_\theta)\Delta y_d^m$. There is no contradiction for those cases where one of these holds but not the other, as the property does not make any statements about behaviour when the conditions, e.g. $\Delta x_{\bar{d}}^{\bar{m}} > \Delta x_d^m$, do not hold. Due to the two sub-properties holding for different payoff combinations, the union of the lie regions will contain a kink at the intersection of the two lie regions described by the two sub-properties which permits that behaviour can vary across lie types (see Figure 1(e)). When Properties 3(1) - (3) hold simultaneously, the lie region either coincides with that for the case when sub-properties (1) and (2) hold simultaneously or it will exhibit a kink at each of the two intersections.

Figure 1 illustrates the implications of Property 3 for an example. In the figure, the x-axis displays the payoff gain from reporting an untrue state s_d^1 relative to the true state s_d^0 for the DM, $x_d^1 - x_d^0$. The y-axis displays the equivalent for the partner, $y_d^1 - y_d^0$. This means that effectively, the payoffs from telling the truth are normalised to $(0, 0)$. For that reason, every dot represents a binary choice problem: it displays the monetary payoffs of lying relative to telling the truth, the normalised payoffs from telling the truth as well as the decision to lie. An empty circle indicates that we have not observed any choices for that decision problem. A red cross indicates that we have observed that a DM reported the lie for that choice problem. In the example, we have observed that the decision-maker

reported a lie when the relative payoffs were $\Delta x_d = 2$ and $\Delta y_d = 2$. Each sub-figure displays the lie region, indicated by the grey area, depending on which sub-property is assumed to hold. Recall that the lie region gives the set of decision-problems for which we anticipate that the DM will lie based on having observed that he lied for decision-problem d. Sub-figures (d) - (f) show the implications when more than one sub-property hold simultaneously.

The lie regions themselves are a valuable tool to predict behaviour based on having observed a DM's choices for a limited number of decision-problems. The tool can easily visualise how a DM is expected to behave for decision-problems that were not observed but that fall within the lie region. The lie regions thus provide the researcher with a tool to help anticipate behaviour without requiring additional data. At the same time, it can be used to falsify the model.

3. Experimental Design and Logistics

The experiment was designed with three requirements in mind. The first was to ensure that lying decisions are observable at the individual level without confounding beliefs in order to elicit lying preferences and assess the role of different types of lies. The second was the identification of types of decision-makers and the distribution of types in the data. Third, the design had to permit the clean separation of lying preferences from social preferences.

This section discusses first the logistics and then the design of each element of the experiment in detail.

3.1. Logistics

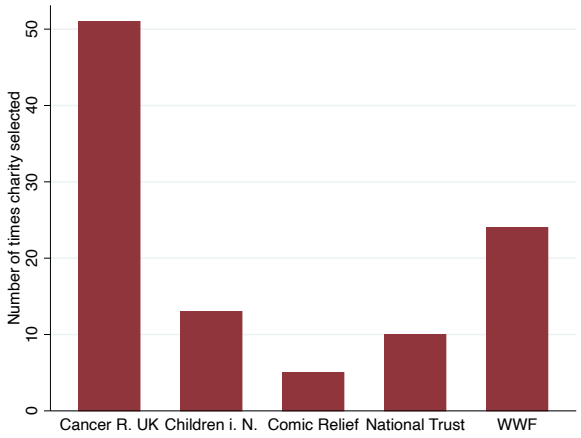
The experiment was conducted online with subjects recruited via the experimental platform Prolific.[1] In total, 103 subjects (51.5% females, mean age was 36.6), which corresponds to roughly 88% of completed responses, passed the comprehension checks. Subjects spent on average 25 minutes on the experiment.

At the beginning of the experiment, subjects were informed that they would be matched with a charity that they could select from a list of five well-known and popular UK charities. The selected charity plays the role of "partner" throughout the experiment. The rationale for selecting a charity as partner is explained in detail in the subsection below. The majority of subjects who sign up to Prolific are from the the UK which is why UK charities were selected. While these charities are very popular in the UK, it is unlikely

[1]Peer et al. (2017) and Palan and Schitter (2018) document the high quality of this platform relative to alternatives, both with respect to the participants and the functionality of the platform.

that someone with no ties to the UK is aware of their existence. For that reason, the subject pool was restricted to UK nationals. The five charities were: Cancer Research UK, Children in Need, Comic Relief, National Trust and World Wildlife Fund. These charities are active in the following areas respectively: medical research, child welfare, poverty relief, cultural heritage and wildlife support. The broad range of charities was selected to increase the probability that subjects cared about at least one of these charities. The charities were paid in the form of donations. Figure 2 shows how many subjects chose each of the charities. After subjects had selected their most preferred charity, they played the main experiment.

Figure 2: Choice of Charities



The figure shows the number of subjects that chose each charity. The charities' names in full are: Cancer Research UK, Children in Need, Comic Relief, National Trust and World Wildlife Fund.

The experiment consisted of two main stages, a lying and a social preference game. Subjects were randomly allocated to starting with one of the games and then played the remaining game. After having completed the main stages, they were then asked to respond to a questionnaire that included demographic questions, a cognitive reflection test (CRT) and a Big 5 test. The CRT test contained four questions; three questions from a recent version of the test (Thomson and Oppenheimer (2016)) and one from the traditional CRT test (Frederick (2005)). The majority of questions was chosen from the recent version in order to reduce the likelihood that subjects already knew the answers by heart. This may be the case for questions from the traditional CRT test as subjects often encounter the traditional test in online experiments and may thus be familiar with the correct responses.[2] The question from the traditional CRT test was included to ensure comparability. The Big 5 test used was a short, ten item version (Rammstedt and John (2007)) in order to

[2]At the time the experiment was conducted, answers to the recent version of the CRT test were difficult to find online, thus reducing the chance of cheating.

reduce the time spent on the experiment.

In addition to a show-up fee, subjects were paid based on their choices in two rounds of the experiment, one from each of the main stages. The random selection of one round for payment from each stage should reduce the likelihood of balancing behaviour with respect to payments to the charities. The rounds were selected by a random number generator. On average, subjects were paid 7.61 GBP.

3.2. Design

3.2.1. Lying game

Subjects played 60 rounds of a lying game developed for this paper, after having completed a game specific comprehension check. This game will be referred to as "the lying game". The lying game is informed by the framework introduced in Section 2. As such, each subject takes the place of a decision-maker who is matched with a passive partner, the charity, and each round forms one decision-problem d. The terms "round" and "decision-problem" will be used interchangeably from here on. In each round, a true state is privately observed and the DM is asked to report this true state but has the opportunity to lie.

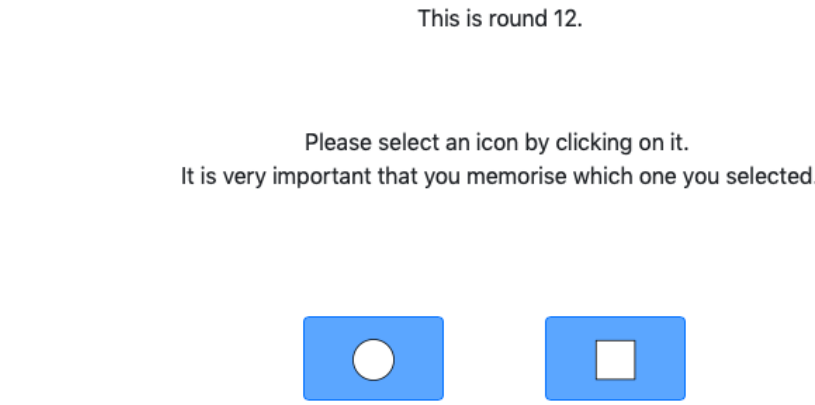
In the framework, subjects' choices affect a partner, the charity, who is passive.[3] Using a charity rather than another player as partner has several advantages in creating such a setting. First, the charity does not participate in the experiment and it is thus obvious to the subjects that their partner is passive which ensures that there are no first or second order beliefs about the partner's behaviour. Second, subjects actively select the charity which ensures that the partner is salient. Third, choosing from a list of several charities with different purposes ensures that they care, at least to some extent, about this partner. For control, I ask subjects to self-report how much they care about the charity that they have selected and charitable giving in general (82.52% reported that they liked the charity that they had chosen either moderately, very or extremely).

Each of the rounds consisted of two screens that the subjects saw sequentially. On the first screen, geometric shapes were displayed on buttons (for an example see Figure 3). These shapes consisted of easily recognisable geometric shapes, such as circles, squares, triangles, pentagons, diamonds and hexagons. Subjects were asked to choose one of the shapes by clicking on it and were told to memorise that geometric shape.

On the second screen, subjects saw the same shapes on buttons but now they were associated with monetary payoffs. Specifically, a table displayed the payoffs associated

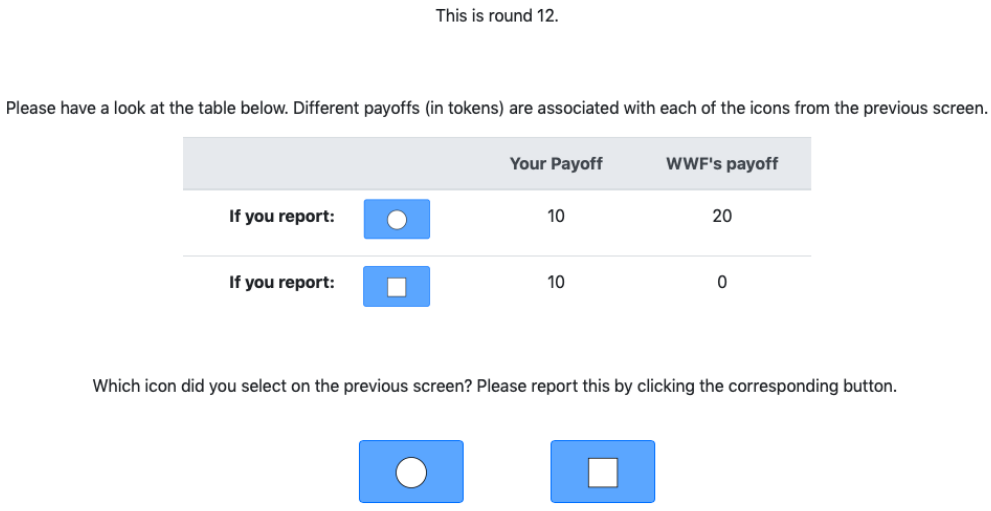
[3]Subjects selected the charity before they knew the specifics of the games or that it was a lying game.

Figure 3: Example of the 1st screen of one of the rounds of the lying game



with each of the geometric shapes from the first screen (see Figure 4). Each shape thus corresponded to one state s_d^m in the framework. The payoff table showed one payoff for the subject, x_d^m, and one for the charity, y_d^m, for each of the shapes. Payoffs were given as an experimental currency (EC), subjects were paid in GBPs and were told the conversion rate of 5 tokens = 1 GBP in the instructions. In order to increase the salience of the partner, the name of the charity that the subject had chosen was displayed in the table.

Figure 4: Example of the 2nd screen of one of the rounds of the lying game



At the bottom of the screen, subjects were asked to report which geometric shape they had chosen on the first screen. The choice on the first screen thus determined the true state s_d^0. If the subject reported the same shape on the second screen as he had clicked on in the first, his report was classified as telling the truth. If, however, he reported another state, the report was classified as a lie. As the experiment was conducted online and the

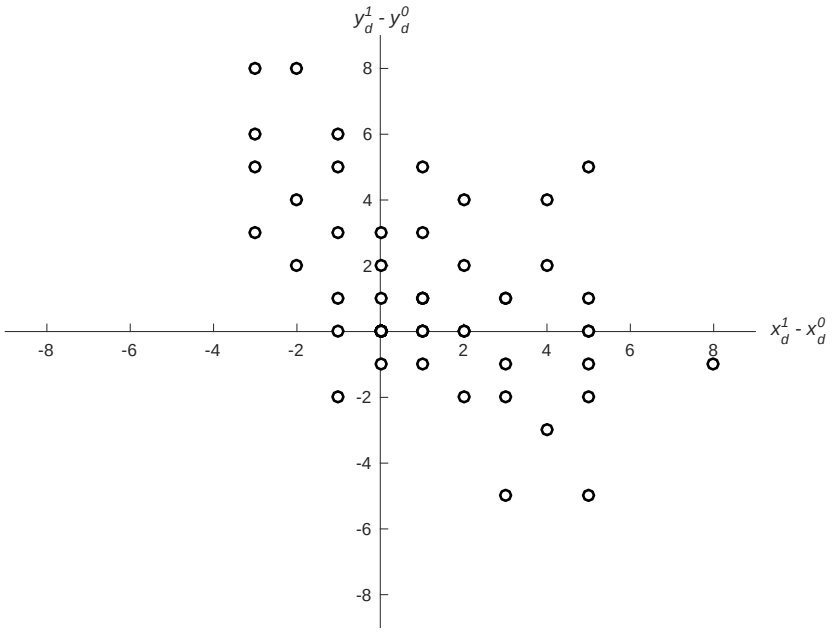
initial choice as well the report were registered by the software, individual lying behaviour was observable to the researcher. Importantly, in conjunction with the use of the charity as partner, this implies that choices are observable at the individual level but do not suffer from (higher order) beliefs stemming from the expected reaction of the partner to the report. The type of the potential lie could be varied across rounds by changing the payoffs relative to the selected shape i.e. true state.

To illustrate, imagine that a subject is currently in round 12 of the lying game, $d = 12$. On the first screen of round 12, the subject has the choice between choosing a circle or a square (as in Figure 3). Further, imagine that he chooses the square. The true state of the twelfth round s^0_{12} is then "square" and there exists one untrue state s^1_{12} which is "circle". On the second screen, he is informed that the square is associated with 10 ECs for himself and 0 ECs for the charity while the circle is associated with 10 ECs for himself and 20 ECs for the charity (see Figure 4). Reporting the circle would constitute telling a *weakly altruistic lie (WAL)* and reporting the square would constitute reporting the true state, s^0_d. If, instead, the payoffs associated with the circle had been 20 ECs for both, reporting the circle would have constituted a *mutually beneficial lie (MBL)*. The geometric shapes themselves were randomised across rounds and subjects to prevent shape-based effects. The order of the 60 rounds was randomised across subjects as well, to reduce order effects. To prevent that subjects simply clicked through the rounds rather than making choices based on their lying preferences, buttons changed positions between rounds and between screens within rounds so that it was impossible for someone to keep the mouse cursor in the same position and then click through the whole experiment.

All subjects played the same 60 rounds. Rounds differed with respect to the particular payoffs and the number of available choices. In 50 rounds, DMs were faced with a binary choice between telling the truth, reporting s^0_d, and telling a lie, reporting s^1_d. In ten rounds, more than one lie type was available for reporting in order to be able to directly research preferences between lie types. Here, DMs were given a choice between telling the truth and telling one of multiple lies rather than a binary choice between telling the truth and telling a lie. The specification of each round are given in Table 5 in the Appendix.

The framework allows for heterogeneity in the utility function and one of the goals of the experiment is to identify whether and where such heterogeneity exists. Therefore, the payoffs across rounds were systematically selected to maximise variation in behaviour across the lie type space. Several pilots examined behaviour for a wide range of payoffs. To tease out the heterogeneity, more rounds were located close to payoff combinations where the differences across subjects were expected to become apparent, based on results from the pilots, and where a fine payoff grid was thus necessary. Figure 5 shows the space of

Figure 5: Payoff space for binary questions



All rounds with two states (true state s_d^0 and an untrue state s_d^1). The horizontal axis displays $x_d^1 - x_d^0$ and the vertical axis displays $y_d^1 - y_d^0$. The upper right quadrant thus displays *mutually beneficial lies*, the lower right quadrant displays *egoistic lies*, the lower left shows the *mutually harmful lies* and the upper left shows the *altruistic lies*. Weak types, for example *weakly altruistic lies*, are displayed on the axes. Some rounds use the same payoff differences but have different levels of payoffs so that some dots describe multiple rounds.

payoffs for rounds with two states, s_d^0 and s_d^1. The x-axis displays the change in the DM's payoff from reporting s_d^1 compared to reporting the true state s_d^0, and the y-axis displays the same for the charity's payoff. Each dot represents at least one round. To control for possible level and inequality effects, some payoff differences were used multiple times so that multiple rounds, with different absolute consequences but identical relative ones, can be represented by one dot. For example, imagine that in one round d with two states, the true state's payoffs are $(x_d^0, y_d^0) = (5, 5)$ and the untrue state's payoffs are $(x_d^1, y_d^1) = (3, 7)$, while in another round $\bar{d}$ the payoffs are $(x_{\bar{d}}^0, y_{\bar{d}}^0) = (7, 3)$ and $(x_{\bar{d}}^1, y_{\bar{d}}^1) = (5, 5)$. In both rounds, the payoff differences are $x_d^1 - x_d^0 = x_{\bar{d}}^1 - x_{\bar{d}}^0 = -2$ and $y_d^1 - y_d^0 = y_{\bar{d}}^1 - y_{\bar{d}}^0 = 2$. Both rounds are therefore visualised by the same dot in Figure 5. However, in the second round, lying generates equality of payoffs while in the first round, lying generates inequality. These inequality concerns are likely to affect lying behaviour. To control for such considerations, the payoffs for the rounds have been selected to ensure that each lie type is represented by situations that vary with respect to whether a potential lie reduces or increases inequality.

Behaviour in the experiment can be treated as revealed preferences when the framework

is applied to the experiment, as explained in Section 2. Specifically, the identifying assumptions, Properties 1 and 2, imply that any untrue state that is reported reveals a strict preference for lying in that round and thus allow to map behaviour into preferences. It is then possible to test whether there exist subjects who always choose to report the true state s_d^0 and who can be classified as never-liars and whether some subjects lie occasionally. The reported states and the attached payoffs can be used to examine whether there exists a payoff combination for which a subject switches from telling the truth to telling a lie. Based on these inflection points, subjects can be classified into several types of DMs, θ. Finally, behaviour in the experiment can be assessed on whether changes between lying and telling the truth are systematic and if yes, whether the regions of inference property, Property 3, can describe behaviour.

The design introduced above was selected with the aim to isolate lying preferences. A possible concern with the lying game could be that when subjects misreport which icon they had selected, this might be due to memory issues. However, subjects had to memorise only one geometric shape per round for a few seconds and the shapes were easy to remember. Systematic and frequent choice errors are thus unlikely. In addition, more than a third of subjects did not lie at all which suggests that a large fraction of subjects was able to correctly remember the choice they had made on screen one in each of the rounds. This supports the claim that it was easy to remember the initial icon choice. Moreover, a large majority of subjects was very systematic in their behaviour, reducing the likelihood of memory errors as primary driver and further illustrating the success of the design in capturing behaviour across lie types.

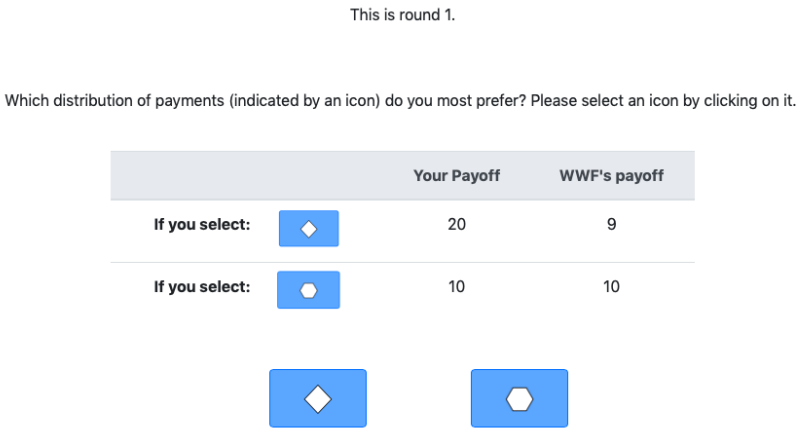
3.2.2. Social preference game

When examining lying behaviour in settings with two individuals, a key point of interest is whether subjects' responses in the lying game perfectly, or highly, correlate with their social preferences and are thus not so much driven by lying considerations but by their social preferences. For this reason, the experiment contained a social preference game which has been designed to address this concern directly.

The social preference game was thus designed to be directly comparable to the lying game. Consequently, this game also consisted of 60 rounds. Each round was identical to the corresponding round from the lying game with respect to the choice sets. However, to elicit social preferences, there was one exception: subjects were only presented with the second screen which displayed the payoff table. Here, they were asked to select their most preferred option. The key difference is that because subjects only see the second screen, there is no truth benchmark for the round any more. Subjects are free to choose their

most preferred payoff bundle without any concerns about lying or truth-telling entering the decision-making process. This allows the researcher to directly contrast choices in a distributive setting to those in a lying setting. Figure 6 displays an example screen from one round of the social preference game.

Figure 6: Example screen of one of the rounds of the social preference game



4. Results of the Lying Game

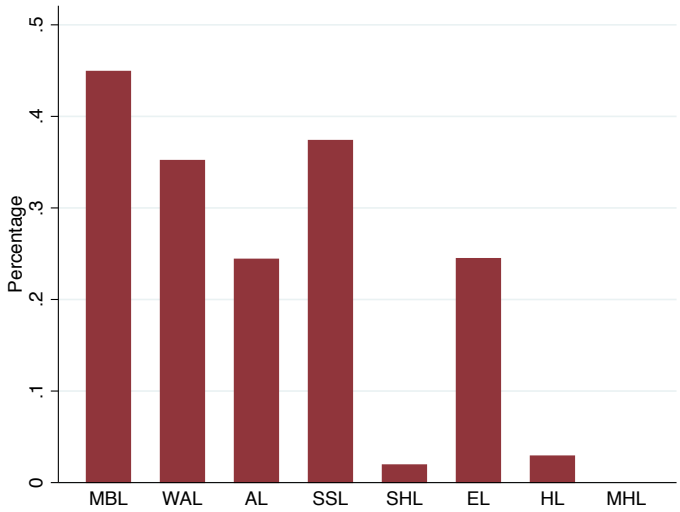
This section presents the results of the lying game. It starts by presenting the aggregate experimental results to facilitate comparison with the literature. It then examines whether lie types matter, measures the heterogeneity of decision-makers and the distribution thereof and examines whether the properties hold in the data. The relationship between social preferences and lying preferences in this experiment is discussed in the section that follows.

4.1. Aggregate results

In the experiment, 31.11% of subjects lied on average per question. This percentage varies drastically across lie types, demonstrating the importance of subdividing lies into types. Figure 7 shows the mean percentage of lies told across the lie types. The percentage of lies averaged by lie type is given by the red bars. They show that the highest percentages of lies, 35% to 45%, occurred for lie types where at least one out of the two subjects benefits from the lie without the other suffering, namely *mutually beneficial*, *weakly altruistic* and *self-serving lies*. *Altruistic* and *egoistic lies* have similar rates of lying of around 25%. Very few lies were told that harm either the DM or the charity relative to the payoff connected to the truth, i.e. *self-harming*, *harmful* or *mutually harmful lies*. This would have been

different if another context had been provided, for example in-group out-group effects, where it would have been possible that a DM would have been willing to incur a small loss to severely harm someone from an out-group. Differences in behaviour across chosen charities are shown in the Appendix.

Figure 7: Percentage of liars by lie type



The percentage of liars averaged by lie type is given by each bar. The acronyms stand for *mutually beneficial lie* (MBL), *weakly altruistic lie* (WAL), *altruistic lie* (AL), *self-serving lie* (SSL), textitself-harming lie (SHL), *egoistic lie* (EL), *harmful lie* (HL) and *mutually harmful lie* (MHL).

Next, I compare the percentages of lies told in the literature to those found in the lying game. In order to be able to compare the results across lie types, my results need to be compared with several papers as most of them examine a small number of lie types, often one, per paper. The results are displayed in Table 2. As it is a seminal paper, the first column shows the results of Gneezy (2005). I also report the results of Hurkens and Kartik (2009) as they supplement Gneezy (2005)'s findings by increasing the variation of lies considered in the paper. To obtain a sense of percentages of lies told for lies that are not egoistic, I also include Biziou-van Pol et al. (2015) and Erat and Gneezy (2012).

The table reveals that there is substantial variation in the percentage of lies, between as well as within lie types. For example, to see the variation between lie types, Biziou-van Pol et al. (2015) find that 83% of subjects told a *mutually beneficial lie* compared to 23% who told an *altruistic lie*. To see the variation within lie types across types, compare Erat and Gneezy (2012)'s finding that between 49% and 65% of subjects told a MBL with the 83% from Biziou-van Pol et al. (2015). The variation suggests that the percentage of lies depends both on the specific payoffs linked to a question and on the lie type. It is thus important to include many specifications of a lie type that differ with respect to

Table 2: Percentage of liars by lie type: Comparison with literature

Lie Type	Gneezy (2005)	Hurkens & Kartik (2009)	Biziou-van-Pol et al. (2013)	Erat & Gneezy (2012)	This paper
EL	17 - 36%	38 - 47%	NA	37%	12 - 40%
AL	NA	NA	23%	33%	4 - 35%
MBL	NA	NA	83%	49 - 65%	40 - 50%
SSL	NA	NA	NA	52%	32 - 41%
WAL	NA	NA	NA	NA	32 - 42%
SHL	NA	NA	NA	NA	2%
HL	NA	NA	NA	NA	3%
MHL	NA	NA	NA	NA	0%

Example percentages of lies told per lie type in the literature compared to this paper. The acronyms stand for *egoistic lie* (EL), *altruistic lie* (AL), *mutually beneficial lie* (MBL), *self-serving lie* (SSL), *weakly altruistic lie* (WAL), *self-harming lie* (SHL), *harmful lie* (HL) and *mutually harmful lie* (MHL). NA signifies that a paper did not include the lie type listed in the row. When a paper examined at least two instances of a lie type, the range of percentages of lies told by subjects is given.

the relative payoffs, equality concerns, level effects etc., for each lie type in an experiment. Note that the ranges of percentages of lies told by lie type are compatible with the results of the papers listed. For an analysis of the relationship between percentages of lies told and covariates such as cognitive reflection test or Big 5 scores, see the Appendix.

4.2. Do lie types matter empirically?

The previous section has provided evidence that people behave differently across lie types. This section formally analyses the degree to which lie types matter and how they should be included in lying frameworks.

To this end, I examine the degree to which two baseline utility models from the literature can explain the experimental data as a whole. I then compare their performance to two models that differ with respect to the role of the lie type in the utility. The first model allows the constant cost of lying, c_θ, to vary across lie types while the second model additionally allows the lie type to affect the utility from the payoffs of a lie i.e. where the lie type enters $v(\cdot)$. Comparing the baseline models to the augmented models can tell us whether lie types have a significant impact on preferences while comparing the augmented models with each other can tell us how lie types enter preferences.

The baseline model has been selected to fit with the most basic theory of lying: when deciding whether to report the true state s^0 or an untrue state $s^m, m \neq 0$, the DM compares only his own payoff from lying to his payoff when telling the truth but faces a constant cost of lying. The DM's utility from reporting the untrue state s^m, with the utility

from telling the truth normalised to zero, is then given by the utility from the monetary payoff from lying compared to that from reporting the true state minus a constant cost: $u(s^m; s^0, \theta) = \beta_\theta[x^m - x^0] - \mathbb{1}_{s^m \neq s^0} c_\theta$. Papers such as Gibson et al. (2013) use models in this spirit as a baseline. As most papers acknowledge that the DM might also care about a person other than the DM who is affected by the lie, I also use an extended baseline model. In the extended model, the DM's utility depends on both his own payoff as well as that of the partner and contains a constant cost: $u(s^m; s^0, \theta) = \beta_\theta[x^m - x^0] + \gamma\theta[y^m - y^0] - \mathbb{1}_{s^m \neq s^0} c_\theta$. To report an untrue or the true state in such settings is a binary decision and is therefore modelled by logistic regressions. The baseline models' equations are given by:

$$\text{Baseline model} \ : E[Y_\theta | x^0, x^m, \theta] = \frac{1}{1 + exp^{-(\beta_\theta(x^m - x^0) + \tilde{c}_\theta)}} \tag{1}$$

$$\text{Extended baseline model} \ : E[Y_\theta | x^0, x^m, y^0, y^m, \theta] = \frac{1}{1 + exp^{-(\beta_\theta(x^m - x^0) + \gamma_\theta(y^m - y^0) + \tilde{c}_\theta)}} \tag{2}$$

where $Y_\theta = 1$ if a DM of type θ lied, $c_\theta = -\tilde{c}_\theta$ is a constant cost of lying and β_θ and γ_θ are weights on the monetary payoffs.

To analyse how well the models can explain the data, I perform an in sample prediction exercise where the models were fitted to the data using logistic regressions. For the prediction, I use the whole sample to fit the data. The performance of the models is evaluated based on the size of the mean squared forecast error (MSE).

To assess the performance of the model with varying constant costs, I add dummies for the lie types to the extended benchmark model. This reduces the MSE to 0.1969 (min SE of = 0.0004 and max SE = 0.9615). The improvement in the MSE, compared the benchmark and the extended benchmark models, is marginal. This suggests that lie types do not affect lying preferences via shifts in the constant cost of lying.

I then examine the explanatory power of the model that permits lie types to affect the utility from the payoffs, $v(\cdot)$. To capture this, I split the data by lie type and fit the model to each lie type separately. In order to compare the the performance of the lie type specific model to that of the other models, MSEs are averaged across the lie types.

Table 3 displays the MSE of each model as well as the minimum and maximum squared error for both prediction exercises. It shows that the benchmark and the extended benchmark models have a similar performance to each other. If we simply used those models, this would suggest that the partner's payoff plays only a limited role for the decision to lie. When we allow for the constant cost to vary across lie types, there is a very slight improvement in the MSE which suggests that lie types do not impact lying preferences via the constant cost of lying. However, when we examine the MSE of the model that

allows for variation in the utility of payoffs across lie types, there is an improvement of 21.5% in the MSE. These results show first, that lie types play a significant role for lying preferences and second that they enter these preferences by interacting with the utility from monetary payoffs rather than via shifts in constant costs.

Table 3: In sample prediction results

	MSE	min SE	max SE
Baseline model	0.2108	0.0409	0.6366
Extended baseline model	0.2036	0.0146	0.7727
Varying constant cost model	0.1969	0.0004	0.9615
Varying utility from payoffs model	0.1590	0.0190	0.2470

In sample prediction results for the four models considered. Performance of the models is given by the mean squared error (MSE) of the forecast as well minimum and maximum squared errors (min SE and max SE, respectively). Row "Baseline model" displays the performance of the model that only includes the decion-maker's payoff. Row "Extended baseline model" displays the performance of the model that also includes the partner's payoff. Row "Varying constant cost model" displays the performance of a model that includes the decision-maker's and the partner's payoff and allows the constant cost to vary across lie types. Row "Varying utility from payoffs model" displays the performance of a model that includes the decision-maker's and the partner's payoff and allows the utility of the payoffs to vary across lie types.

4.3. Heterogeneity of preferences across individuals

My theoretical framework assumes that individual lying preferences are heterogeneous and the inference region property, Property 3, imposes structure on how this heterogeneity can look like. Specifically, the property suggests that there exist around six groups of decision-makers and lays down expected behaviour for these groups. This subsection examines whether the property can capture subjects' behaviour in the experiment.

To systematically and objectively examine subjects' heterogeneity, I employ an unsupervised machine learning algorithm which creates a partition of subjects into groups.[4] Specifically, I use a k-means clustering analysis which is a statistical pattern recognition tool that, in economics, is commonly used in time-series econometrics (see for example Falat and Pancikova (2015), Bagnall et al. (2003), Focardi and Intertek Group (2001-2004)). k-means is one of the most popular classification algorithms.[5]

[4]While the framework provides a prior for the number of groups and their behaviour, this section will approach the issue objectively, i.e. without imposing these priors, to prevent that the priors bias results in favour of the framework.

[5]Many papers in economics use a related method called mixture models. Mixture models are a widely used tool in economic analysis, primarily to detect and model heterogeneity (see for example Cameron and

The algorithm works in the following way: First, the researcher specifies the number of clusters k and initialises the location of the centroids of the clusters randomly (a glossary of the terminology is provided in the Appendix). All subjects are then categorised as belonging to one of the clusters. A subject is allocated to the cluster with the centroid that has the shortest Euclidean distance to that subject. The algorithm then determines new centroids of the clusters based on shifting the centroid to minimise the distance to all members that have been allocated to it. Then, members are reallocated to the clusters based on the distance to the new centroids. This is repeated until the clusters are "stable": updating the centroids does not affect group membership or the position of the centroids. The method is unsupervised in the sense that the true group membership is unknown to the researcher.

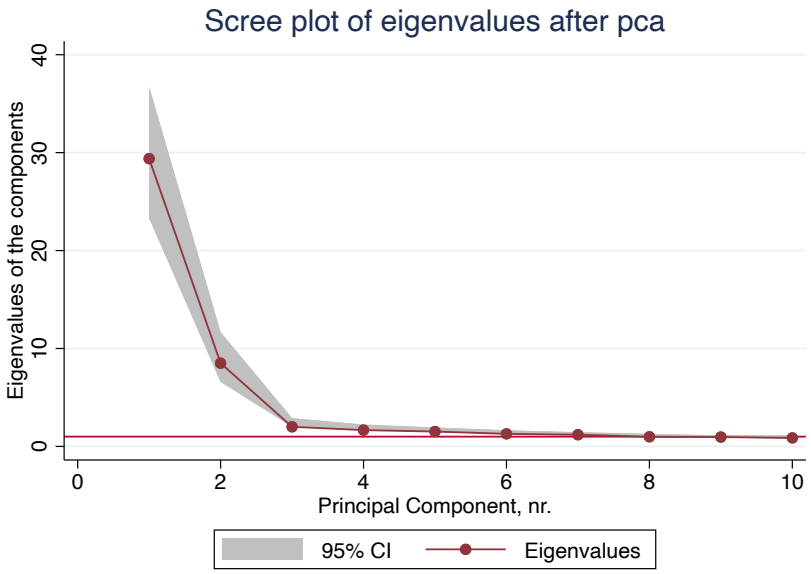
The algorithm is suitable to situations such as identifying heterogeneity of preferences for the following reason: the algorithm requires to know only the number of groups it is looking for and the variables on which it bases the group selection. Furthermore, if the number of heterogeneous groups is unknown, the method can be combined with information criteria with which the number of groups can be selected. The advantage of using a k-means cluster analysis over alternative methods, such as Bayesian methods, is thus that results do not depend on an informative prior beyond knowledge of the variable in which heterogeneity is expected.

In order to weight all variables evenly, variables are standardised by subtracting the mean and dividing the variables by their standard deviation. I then perform a principal components analysis (PCA) on the binary decision to lie of each of the 60 questions and retain only the main principal components, also called factors. PCA is conducted to prevent that clusters are biased towards variables that explain little of the data. Specifically, PCA allows the researcher to reduce the number of variables while preserving the informational content of the 60 lying decisions made by each subject. This procedure is also known as factor model analysis. Factor models are widely used in economics and finance (see for example Engle and Watson (1981), Chamberlain (1983), Stock and Watson (2005)), typically to reduce the number of parameters that have to be estimated. To identify the number of components that should be retained, a scree plot of the eigenvalues after PCA is created. A scree plot is a figure which plots the eigenvalues of the principal components. One then looks for the so called "elbow point" where the information gain of adding another component levels off. Figure 8 shows that only the first three components have

Heckman (1998)), such as identifying utility function shapes across heterogeneous individuals. K-means "is closely related to the EM algorithm for estimating a certain Gaussian mixture model" (p. 510, Hastie et al. (2009)). Specifically, mixture models make a probabilistic assignment of observations to the groups while the k-means algorithm uses deterministic assignments. When the variance of the Gaussian density becomes zero, the two methods coincide (Hastie et al. (2009)). As I am interested in deterministic group assignments, k-means is the preferred method, here.

notable explanatory power and I therefore only include these three components in the k-means analysis.

Figure 8: Scree plot of the 10 largest components in lying game PCA analysis



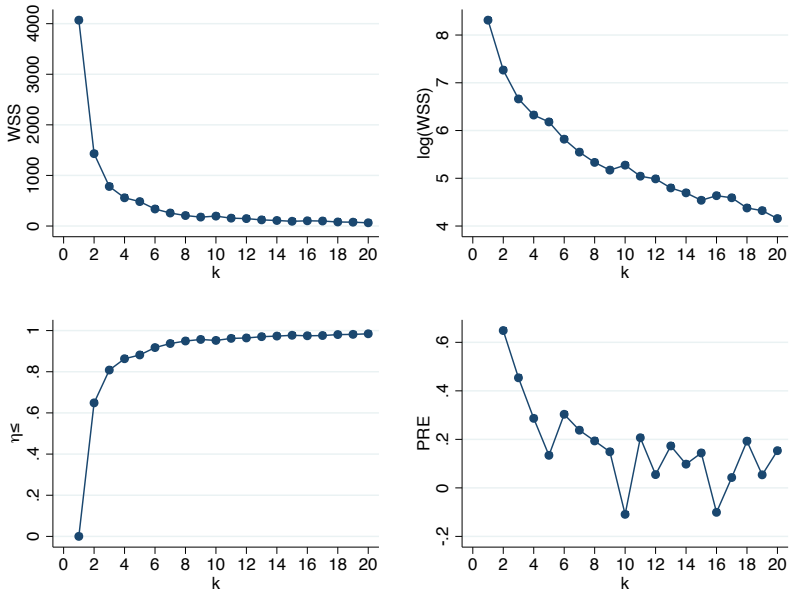
The x-axis shows the largest components (descending) and the y-axis the size of the eigenvalues.

Based on the regions of inference property of the framework, we expect a number of types, and thus clusters, close to six (see the possible sub-property combinations for Property 3). To objectively assess the number of types of DMs in the data, I conducted an initial analysis of how many groups are ideal for the analysis. This consists of repeating the k-means exercise for $k = 1, ..., N$ groups (here, N=20) and compare the total sum of squares (TSS, calculated when $k = 1$) to the within sum of squares (WSS). The optimal number of groups can be found where the improvements in explanatory power are levelling off. Figure 9 shows the performance of the different numbers of clusters. The figure indicates that the number of clusters k should be set equal to six as the informational gains of adding another group level off at around $k = 6$.

The clusters obtained via the k-means analysis with $k = 6$ can explain circa 91% of the variation in the data ($\eta^2 = 0.91$; η^2 is a goodness of fit analysis similar to R^2). We can thus say that behaviour across individuals is highly heterogeneous. The heterogeneous groups of DMs are represented graphically by Figure 10. It shows the binary choice between lying and telling the truth across all questions with a binary choice (one lie and one truth available). Each panel in the figure represents one of the six clusters.

The group easiest to identify, both visually and statistically, is that of never-liars (see

Figure 9: Performance indicators by cluster number in lying game k-means analysis

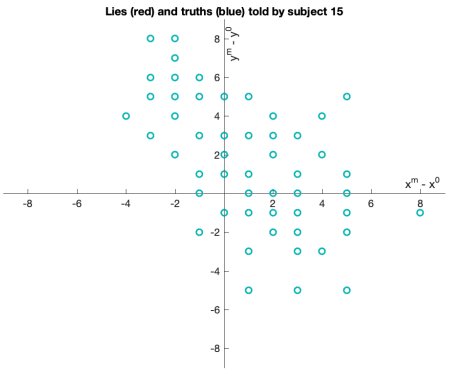


Plot of within sum of squares (WSS), log WSS, η^2 and proportional reduction of error (PRE) for number of clusters $k = \{1, ..., 20\}$. The figures show that gains from adding another cluster level off at $k = 6$.

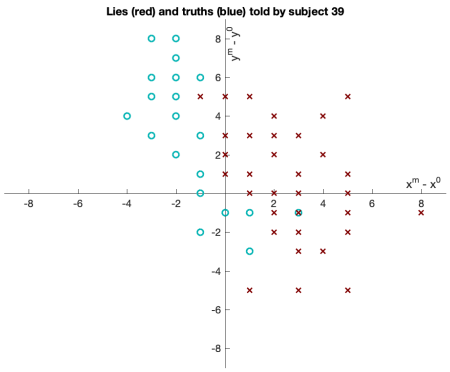
Figure 10(a)). This group contains 44 out of the 103 subjects (some of them misreport a handful of times throughout the experiment but these appear to be errors rather than choices, see Section A.4.2 in the Appendix). Of those, 28 truly never lied corresponding to 27% of the sample. This proportion of never-liars is exactly what we should expect based on the percentages of never-liars found in the literature (for example Erat and Gneezy (2012) find that 35% of subjects don't lie even when this would have resulted in a Pareto improvement in payoffs). The second largest group, with 15% of subjects, is that of subjects who lie whenever there own payoff (weakly) increases, regardless of the effect on the charity (see Figure 10(b)). The third largest group with 14% of subjects is that of subjects who lie both for the gain of the charity and the DM (see Figure 10(c)). Subjects in this group differ with respect to the degree with which they weight the DM's gains/losses relative to that of the charity, corresponding to differences in α_θ. In the fourth group, which also contains 14% of the subjects, are subjects who lie only when the charity gains from the lie (see Figure 10(d)). Subjects in this group show greater within-group variation than those in the previous three groups.

Figure 10: Representative DMs by behavioural lying cluster

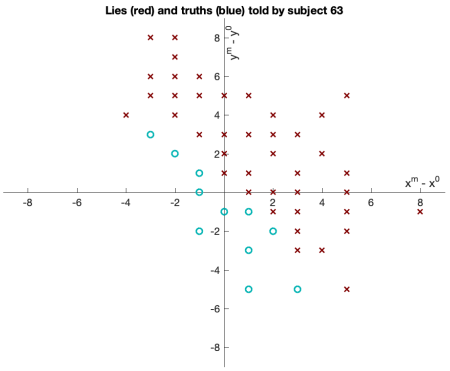
(a) Never-liar



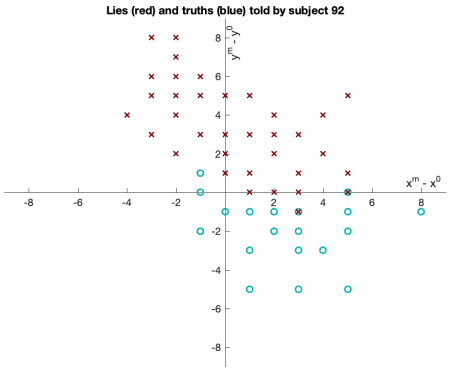
(b) Lies mainly when DM gains



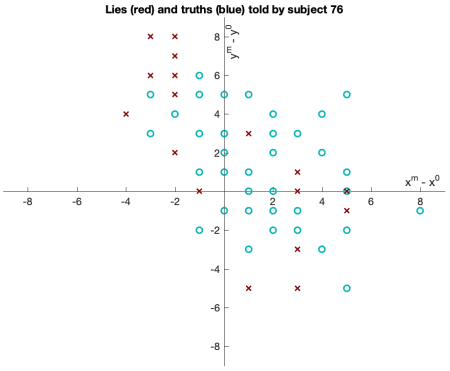
(c) Lies for both DM's and charity's gain



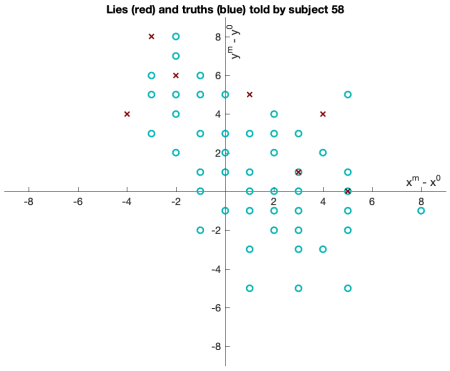
(d) Lies when the charity gains



(e) Balancing?



(f) Non-systematic behaviour



Each panel of the figure shows a representative subject from each identified cluster in the lying game. The x-axis displays $x^1 - x^0$ and the y-axis displays $y^1 - y^0$. The upper right quadrant thus displays *mutually beneficial lies*, the lower right quadrant displays *egoistic lies*, the lower left shows the *mutually harmful lies* and the upper left shows the *altruistic lies*. Weak types, for example *weakly altruistic lies*, are displayed on the axes. Each dot presents an available lie relative to the truth which is standardised to the origin (0,0). Red crosses indicate that the subject lied and a teal dot that they chose the truth instead.

Subjects in groups five and six are similar in so far that their behaviour is non-systematic (see Figure 10(e) & (f)). Together, they account for around 15% of all subjects (16 out of 103). Group 5 contains only two subjects where one of them said that they balanced between helping the charity and helping themselves. It is noticeable that on average, subjects in groups 5 and 6 made more mistakes in the comprehension check of the lying game than subjects from the other groups (43.75% in groups 5 and 6 made at least one mistake compared to 33.4% in the other groups). This may indicate that these subjects did not pay attention to the instructions, providing an explanation of the randomness of behaviour.

Having confirmed that there exist types of DMs who lie in some situations, we can now examine whether the regions of inference property, Property 3, holds for these types of DMs. The property states that once a DM has started to lie for a combination of payoffs defined by the states s^0 and s^m, he should continue to lie for every other decision problem in which a state $s^{\bar{m}}$ has a payoff bundle for which either $\Delta x^{\bar{m}}$ and/or $\Delta y^{\bar{m}}$ is larger than that of the state for which he lied, Δx^m and Δy^m, ceteris paribus. This ensures a kind of monotonicity in the space of Δx and Δy. To examine whether we observe the predicted behaviour in the experiment, we can examine Figure 10. Visually, we should observe red crosses, i.e. lies, only in the upper quadrants and in the lower right quadrant if Property 3 holds, with Property 2 as identifying assumption. To illustrate, imagine that someone lies for a mutually harmful lie. Then, Property 3 implies that he should also tell a neutral lie where the payoffs of the untrue state coincide with those of the true state (as this constitutes an increase in both Δx and Δy and as at least one of (a) - (c) has to hold). However, Property 2 states that for a neutral lie, the DM should always obtain more utility from reporting the true state. We thus have a contradiction. To ensure that both properties always hold, we need to have that DMs do not tell a mutually harmful lie and only start lying when either Δx and/or Δy is larger than zero. Inspecting the figure shows that this is the case for all representative agents across the identified groups. Behaviour in the experiment thus confirms the predicted behaviour from the theoretical framework, Property 3, when assuming that Property 2 holds as identifying assumption.

To show that the six individuals that have been presented in Figure 10 are representative of their clusters, I provide the same figures as above but instead of showing the decisions of an individual, they show what percentage of the subjects of the cluster lied for each question (see Figure 22 in Appendix A.4.3).

Note that I find very systematic behaviour for four of the groups, which indicates that demand or experimenter effects did not play a role for decisions. If the instructions or the set-up of the experiment had biased behaviour in a particular direction, we should

observe that behaviour is biased in one direction. However, the largest group, never-liars, contains the expected number of subjects given the literature so that it does not seem that behaviour was biased in that direction. The other groups that show systematic behaviour are of nearly identical size (15, 15, and 14 subjects respectively). It is highly unlikely that the instructions primed behaviour to follow these three different patterns as well as leading to a representative number of never-liars. Thus, experimenter and demand effects should not be of particular concern.

Finally, I assess how meaningful the identified heterogeneity across subjects is. To this end I conduct two of sample prediction exercises and contrast their performance. For the first exercise, I ignore heterogeneity and for second, I account for it. I can then compare the accuracy of the forecasts and obtain a measure for how much of a difference including heterogeneity makes.

For both exercises, I re-conduct the k-means analysis but this time I only use 40 out of the 60 decision-problems to classify subjects. The 40 questions were selected at random across all lie types; results are robust to varying the 40 questions that are selected. For each subject, I then fit the extended baseline model to the data for the 40 selected questions from all subjects other than the subject of interest. I then predict the behaviour of the subject of interest for those 20 questions that were excluded from the model fitting exercise.

In the first prediction exercise, I do not account for heterogeneity in that I do not discriminate between decision-maker types when fitting the model.

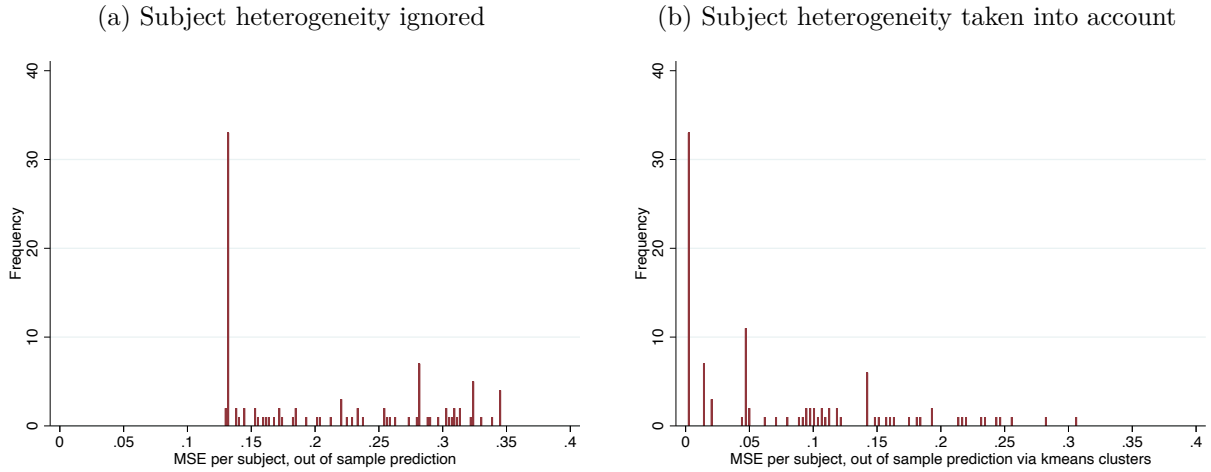
I contrast the performance of the first prediction exercise with that of a prediction exercise which accounts for subject heterogeneity. For this second exercise, I re-conducted the out of sample prediction exercise. The difference to the first prediction exercise is that this time when fitting the model, the data for the 40 questions came from only those subjects who were classified as belonging to the same cluster as the subject of interest.[6] As before, I then use the fitted model to predict the decisions of the subject of interest for those 20 questions that were not included in the model fitting and the k-means exercises.

Figure 11 reports the out of sample forecast performance of both prediction exercises, where prediction accuracy is given by the mean squared forecast error (MSE). Comparing the two panels, we can see that the variance of the MSEs as well as the mean size of the forecasting errors is reduced in the exercise that accounts for heterogeneity compared to the one that does not, panels (b) and (a) respectively. The improvement in the MSEs corresponds to a gain in forecasting accuracy of more than 60 percent. To illustrate, the prediction exercise in panel (a) correctly predicted a subject's behaviour in 79.4% of decision-

[6] As this is an out-of-sample forecast, the individual's behaviour itself is always dropped from the cluster before fitting the model to that group.

problems on average. The out of sample forecasts that did account for heterogeneity, panel (b), in contrast on average correctly predicted behaviour in 92% of the decision-problems. This constitutes a large improvement.

Figure 11: Out of sample performance without and with heterogeneity in preferences taken into account



Histogram of each subject's mean squared error (MSE) in pseudo out of sample prediction exercise. Panel (a) shows that the MSEs are larger when heterogeneity in preferences is ignored during the forecasting exercise compared to panel (b) where it is taken into account.

In order to examine whether the gains from adding heterogeneity to the out of sample prediction exercise are robust the number of clusters specified, I reran the k-means analysis with $k = 5$ and $k = 7$ clusters. The gains from accounting for heterogeneity are robust to slightly changing the number of clusters (see Figure 21 in the Appendix).

In sum, in this section we have learned that lie types enter lying preferences via the utility from payoffs, that subjects are highly heterogeneous and this heterogeneity is both systematic and consistent with that postulated by the framework. Finally, accounting for this heterogeneity in out of sample prediction exercises yields large and significant gains in the accuracy of the forecasted behaviour.

5. Lies and social preferences

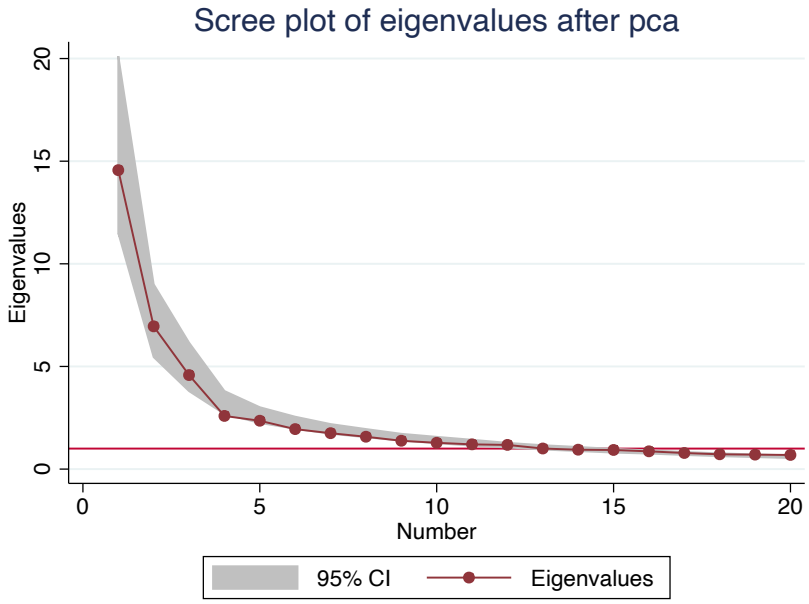
In this section, I examine whether subjects' behaviour changed between the lying and the social preference games and if they did, whether patterns of behaviour changed systematically. The first subsection presents a k-means cluster analysis of behaviour in the social preference game where an algorithm allocates subjects to groups. The group composition is then compared against that of the groups defined by the lying game behaviour. Following the between group analysis, the second subsection analyses the interaction of social and

lie preferences at the individual level.

5.1. Are behavioural groups in the lying game comparable to those in the social preference game?

Section 4 presented a k-means cluster analysis of lie behaviour. Here, I conduct the same analysis but using behaviour from the social preference game, only. Before conducting the k-means analysis, I conduct the same PCA exercise as before, where the 60 decisions are condensed into the principal components. Here, the four largest principal components are needed to capture the variation in the data (see Figure 12). The same seed is used for the initialisation of the clusters and as the aim of this section is to compare the groups from the lie and the social preference behavioural clusters, the number of clusters k is pre-specified to be equal to six.

Figure 12: Scree plot of 20 largest components in social preference game PCA analysis



The x-axis shows the largest components (descending) and the y-axis the size of the eigenvalues.

Figure 13 shows the behaviour of representative individuals for each of the six groups. As with the equivalent figure in the section above, the figures show which alternatives the subject chose for those questions that entailed a binary choice. The origin symbolises the reference point's normalised payoffs, where the reference point is the option that is coded as the truth in the lying game. The x-axis shows the change in the DM's payoff relative to the reference point and the y-axis the corresponding values for the charity. For

example, this means that where both values are positive, picking that alternative would lead to gains for both the DM and the charity relative to the reference point. Red crosses indicate the subject selected that option while teal dots indicate that the subject selected the reference point. I am referring to that option as the reference point only for ease of comparability. If social preferences were the only driver behind behaviour in the lying game, we should expect to the see similar pictures as for the lie groups, except for the never-liar group. If there exists a constant cost of lying, then the patterns of behaviour should be the same but the red crosses should be shifted downward (until the origin) in the social preference game. Of course, it is possible that the same behavioural groups exist but that different members populate them. This will be examined in Subsection 5.2.

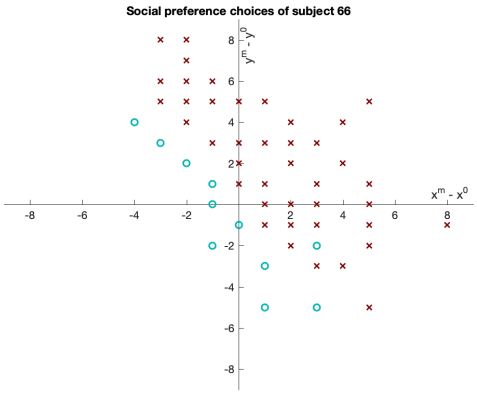
The largest cluster identified by the k-means algorithm contains subjects whose behaviour looks as if they were maximising payoffs of both the DM and the charity with equal, or at least close to equal, weights (see Figure 13(a)). This cluster contains nearly 47% of the subjects. In the second largest cluster, whose members constitute 20.4% of the sample, subjects choose alternatives to maximise the DM's payoff (Figure 13(b)). In the third cluster, which is the same size as the previous one, subjects behave to maximise the charity's payoff (Figure 13(c)). Subjects in cluster four, which correspond to 8.7% of the sample, seem to balance behaviour in so far that they sometimes maximise the charity's and sometimes the DM's payoff and there is no clear pattern (Figure 13(d)). Cluster five contains only three subjects whose behaviour seems to be non-systematic (Figure 13(e)). Finally, cluster six contains only one subject who displays non-systematic behaviour (Figure 13(f)).

The behaviour displayed by the six groups is similar to that identified for the six groups of the lying game. Yet, the size of the respective groups differs between the two games. This suggests that subjects may swap their group membership and thus their behaviour across the games. Here, I assess this possibility at the group level. An analysis at the individual level is provided further below.

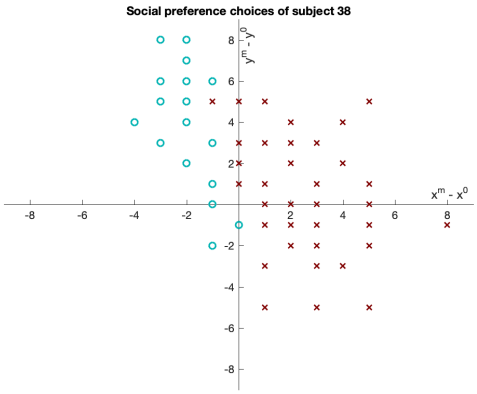
Above, we have seen that the behavioural clusters coincide in the identified behaviour across the lying and the social preference games. However, what we are interested in is not whether the same types of behaviour exist in general but whether the DMs stick to one behaviour across the two games or if they do not, whether the members of the clusters change their behaviour in the same way as the other members of their cluster. To assess this, we need to examine group membership in the clusters. If, for example, those DMs who maximised their own payoff in the lying game are not members of the cluster that maximises the DMs' payoffs in the social preference game, this suggests that DMs change their behaviour across games.

Figure 13: Representative DMs by behavioural social preference cluster

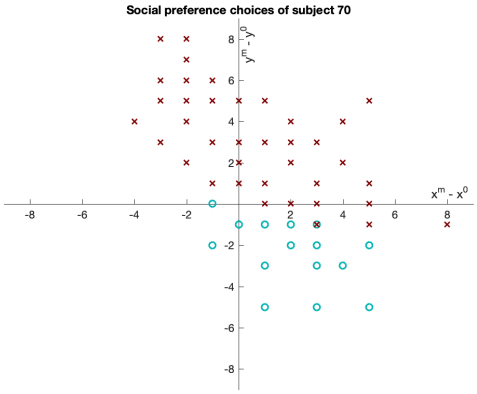
(a) Picks alternatives that maximise DM's and charity's combined payoffs



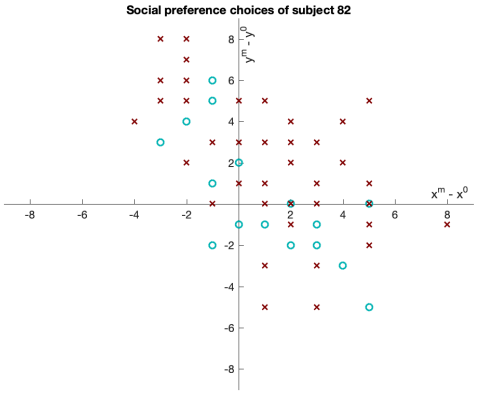
(b) Picks alternatives that maximise DM's payoffs



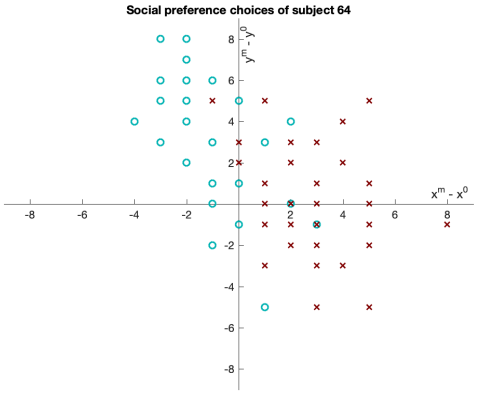
(c) Picks alternatives that maximise charity's payoffs



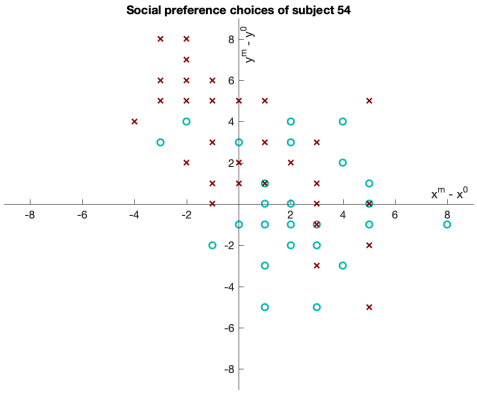
(d) Preferred alternatives from all regions



(e) Choose mainly for DM's sake with more variation



(f) Non-systematic behaviour



Each panel of the figure shows a representative subject from each identified cluster in the social preference game. The x-axis displays $x^1 - x^0$ and the y-axis displays $y^1 - y^0$. A blue circle signifies that the subject chose the "reference point" and a red cross signifies that the subject chose the alternative option. The panels demonstrate high heterogeneity in preferences.

In order to assess whether this is the case, i.e. whether the same people are clustered together in the social preference game as in the lying game, I calculated the normalised mutual information (NMI) score. This score assesses to which percentage group membership coincides across the two games i.e. to which degree the two cluster analyses provide the same, thus mutual, information. The NMI score takes value one if all subjects who have been clustered into one group are also clustered into one group in the second sample. It takes value zero if group membership does not coincide at all. This measure takes into consideration label differences between analyses. For example, if a group of people is clustered into cluster "1" in the lie analysis and the same people are clustered into cluster "2" in the social preference analysis, this measure is able to identify that the people in both clusters are the same, despite being labelled differently. In such a case, the score would be equal to one which would indicate perfect "stability".

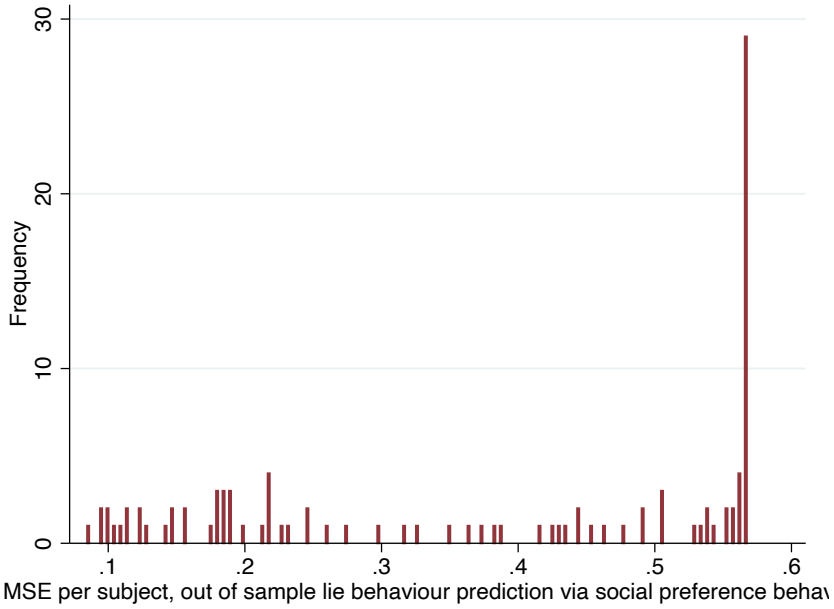
There exist three possibilities. First, DMs could behave in exactly the same way across both games. Second, DMs could change their behaviour but in such a way that most members of a cluster change their behaviour across games in the same way. Third, DMs might change their behaviour but members of a cluster in the the lying game may change their behaviour differently to each other. The first and second case would lead to a very high NMI score as people who have been clustered into one group in the lying game would also be clustered into one group, even if this describes different behaviour, in the social preference game. If, however, the NMI score is low, we would know that the third case is correct.

Running the analysis, I obtain an NMI score equal to 0.238. To compare, if group membership in the clusters from the lying game is compared to fully random clusters, the NMI score is equal to 0.11. If there were perfect correlation between two analyses instead, the score would be equal to 1. A score of 0.238 thus indicates that stability between the two cluster analyses is low and that we are consequently in the third case. This signifies that people who behave similarly in the lie experiment often do not behave similarly to each other in the social preference block and vice versa. Importantly, this implies that social preferences alone cannot explain lie preferences. Otherwise, being in one cluster in the social preference game should have had a higher correlation with behaviour in the lying game.

I further support the finding that lie preferences and social preferences do not seem to be systematically related by showing that one cannot use the behaviour from the social preference block behaviour to predict behaviour in the lie block well. To this end, I run a pseudo-out of sample prediction analysis by individual to account for heterogeneity. I fit the extended benchmark model to the subject's social preference block behaviour and then use this fitted model to predict the subject's behaviour in the lying game. I repeat

this for each subject in the sample. The mean squared forecast error is equal to 0.3834 which is much higher than that in previous analyses. Its size implies that using a subject's social preference block behaviour to forecast their lie block behaviour is only a bit better than using a coin flip to predict their behaviour. Figure 14 shows a histogram of the mean squared forecast errors. Comparing this figure to the results in Figure 11 shows that the forecasting errors here are very large and that social preference game behaviour is thus not suitable to forecasting lying behaviour.

Figure 14: Performance of predicting lying behaviour based on social preference behaviour



Histogram of each subject's mean squared error (MSE) in pseudo out of sample prediction exercise. Behaviour in the social preference game is used to predict behaviour in the lying game.

In summary, while similar types can be classified in the social preference game as in the lying game, they are exhibited by different subjects across the two games. Importantly, subjects who share patterns of behaviour in one of the games mostly do not share patterns of behaviour in the other game. Social preferences thus do not predict lying preferences.

5.2. Individual specific analysis

I now examine how subjects differ between the two blocks and, should there be no patterns across the whole sample, whether there are sub-group specific patterns. To examine these questions, I compare cluster membership in the social preference with that in the lie block subject by subject.

Never-liars, who form the largest lie block cluster, differ a lot from each other in the social preference block. Some of them behave to maximise their own payoff, others that of the charity and many to maximise a combination of the two payoffs. This is of great interest as these subjects acted to fulfil the moral code of "you should not lie". It would have been conceivable that they would act according to another moral code, such as "you should help others irrespective of the effect on yourself", in the social preference game. Instead, never-liars do not seem to share the same preferences in the social preference game.

When we analyse the other lie clusters, other interesting patterns emerge. For the lie cluster that maximises DM's payoffs, roughly two thirds also maximise the DM's payoff in the social preference block. However, the remaining one third behaves more egoistically in the lie block than in the social preference block where they maximise the combination of the DM's and charity's payoffs. For the lie cluster that maximises the combination of the DM's and the charity's payoffs, all subjects behave the same way in the social preference block. This implies that the extended benchmark model with a weakly positive cost of lying is sufficient to describe behaviour for this group of subjects. It also explains why the NMI score was low but larger than if the clusters had been randomly assigned: for this group, cluster membership coincides across the two games. For the lie block cluster that maximises the charity's payoff, a third of subjects also maximise the charity's payoff in the social preference block. The other two thirds act more altruistically in the lie block than in the social preference block where they maximise the combination of the DM's and charity's payoffs. For the two lie block clusters that contain non-systematic behaviour, subjects' social preference block behaviour varies; some maximise the DM's payoff, some the charity's, some the combination of the payoffs and some behave non-systematically.

This reveals three patterns: A third of the subjects is consistent across the two blocks (32% of the whole sample); never-liars (43% of the whole sample) do not share the same social preferences among themselves, and there exist subjects who behave more altruistically and some who behave more egoistically in the lie block than in the social preference block (roughly 13% of the whole sample).[7] A narrative explanation for the latter two could be statements along the line: "If I lie, it should be to benefit someone else" and correspondingly, "If I lie, it should be worth my while and benefit me".

The results show that never-liars do not have common social preferences, and they therefore account for at least some of the variation in the group allocation between the lie and the social preference games. To check whether the low NMI score is driven by these never-liars, I reconducted the k-means analyses without the never-liars. For the analyses,

[7]The remaining 12% stem from subjects who are classified as non-systematic in the lie block and where systematic deviations are therefore difficult to assess.

I used $k = 5$ as one of the six groups identified in the initial k-means analysis of the lie game behaviour was formed of never-liars. Conducting the identical k-means analyses from Sections 4 and 5, including the same seed and PCA pre-step, but having dropped the never-liars beforehand and setting $k = 5$, yielded an NMI score of 0.2872. Recall that this score can be interpreted as the percentage of subjects who shared group membership across the lying and the social preference game, with the score equal to 1 if the groups perfectly coincide across games. This higher score, compared to the score of 0.238 for the analysis with never-liars included, confirms that part of the variation in the groups was driven by the never-liars. However, the score also shows that only around 30% of the remaining subjects share group membership in both games. This confirms that subjects' behaviour varies significantly between the lie and the social preference games.

In summary, both group and individual level analyses confirm that while social preferences enter lying preferences, they cannot fully capture them. The results thus demonstrate a need for theoretical frameworks that analyse lying preferences as something different from social preferences, while taking into account that social preferences matter to some degree, such as the framework introduced in this paper.

6. Parametrised and Calibrated Utility Function

Given that the previous sections have shown that behaviour across lie types is systematic, we can now exploit this for model building and thus for prediction. In this section, I present an example of a parametric utility function that fits the theoretical framework as well as the results from the experiment. This model can be estimated and can be used for model building in a variety of settings. While this is by no means the only possible utility function, it fits the data very well while being relatively simple, given the complex interaction between lie types and decision-maker types. The experimental results sections above have already shown that most subjects behave differently when they are in a distributive scenario without communication as compared to a lie setting with distributional effects. Therefore, the function presented here describes behaviour in the presence of lies, only.

6.1. Utility Function

I propose a utility function of the following shape:

$$u(s_d^m; s_d^0, \theta) = max\{\alpha_\theta(x_d^m - x_d^0) + \beta_\theta(y_d^m - y_d^0), \delta_\theta(x_d^m - x_d^0) + \gamma_\theta(y_d^m - y_d^0)\} - \mathbb{1}_{s_d^m \neq s_d^0} c_\theta \tag{3}$$

Notice that $u(s_d^0; s_d^0, \theta) = 0$ and that therefore if $u(s_d^m; s_d^0, \theta) > 0$, then $r_d \neq s_d^0$ and else $r_d = s_d^0$. The following constraints ensure that the properties imposed by the framework hold:

Constraints:
$c_\theta \geq 0$
If Property 3(1) holds: $1 = \alpha_\theta = \delta_\theta, 0 = \beta_\theta = \gamma_\theta$
If Property 3(2) holds: $0 = \alpha_\theta = \delta_\theta, 1 = \beta_\theta = \gamma_\theta$
If Property 3(3) holds: $1 > \alpha_\theta = \delta_\theta > 0, \beta_\theta = \gamma_\theta = 1 - \alpha_\theta$
If Properties 3(1) & 3(2) hold: $\beta_\theta = 0, \delta_\theta = 0$
If Properties 3(1) & 3(3) hold: $\frac{\delta_\theta}{\gamma_\theta} > 0, \beta_\theta = 0$
If Properties 3(2) & 3(3) hold: $\frac{\alpha_\theta}{\beta_\theta} > 0, \delta_\theta = 0$

Specifically, the constraint on c_θ ensures that Property 2 holds for the utility function. The other constraints ensure that Property 3, the regions of inference property, holds. The properties thus influence strongly which behaviour is allowed within the framework: The properties provide structure on the utility function in the form of constraints.

The utility function consists of a function of the relative payoffs, $v(\cdot)$, and a constant cost parameter. The $max\{\cdot, \cdot\}$ operator permits that behaviour varies across Δx^m and Δy^m and thus across lie types so that $v(\cdot)$ also depends on the lie type. Parameters $\alpha_\theta, \beta_\theta, \delta_\theta, \gamma_\theta$ determine to which degree the decision-maker cares about their own payoff compared to the partner's payoff. If $\alpha_\theta = \delta_\theta$ and $\beta_\theta = \gamma_\theta$, the model collapses to the extended benchmark model that was introduced earlier. Due to the $max\{\cdot, \cdot\}$ operator, the degree to which the decision-maker cares about the payoffs can vary across the payoff space. The approach is similar to piece-wise linearity.

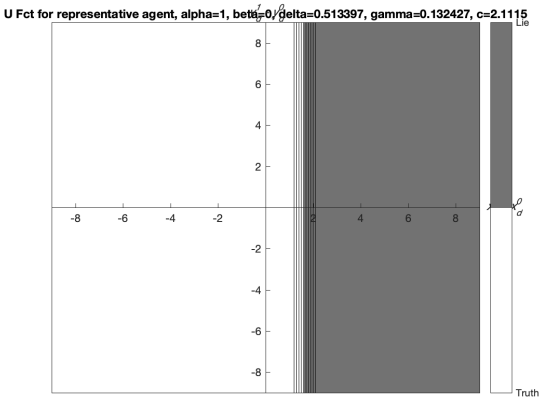
6.2. Calibration

I first adopt a representative agent approach and in a second step demonstrate that we should consider heterogeneity of decision-makers to model behaviour.

First, I calibrate the model to describe the representative agent. To this end, I perform an MLE estimation with a probit likelihood for each of the subjects and then calculate the median estimates. This yields behaviour as shown in Figure 15. The figure shows estimated lie and truth-telling regions (lie regions in grey, truth-telling regions in white) for the representative agent with the gain from lying to the DM shown on the x-axis and that of the charity on the y-axis.

Figure 15 suggests that DMs only care about their own payoffs and start to lie as soon as the payoff gain from lying, $x_d^m - x_d^0$, is larger than a positive constant, the constant cost

Figure 15: Estimated lie regions for the representative agent of the lying game



The x-axis displays $x^1 - x^0$ and the y-axis displays $y^1 - y^0$. The upper right quadrant thus displays *mutually beneficial lies*, the lower right quadrant displays *egoistic lies*, the lower left shows the *mutually harmful lies* and the upper left shows the *altruistic lies*. Weak types, for example *weakly altruistic lies*, are displayed on the axes. Dark grey areas indicate that the agent is expected to lie and white areas indicate payoff combinations for which the agent is expected to tell the truth.

of lying. The figure shows a very distinctive pattern of behaviour. However, the previous sections have shown that lying behaviour is highly heterogeneous. I therefore repeat the exercise for each of the clusters from Section 4. I thus obtain six representative agents whose behaviour can be compared.

This methodology has the advantage that the MLE estimates themselves are not conditioned on the clusters as the model is calibrated to each DM individually. Therefore, no prior on group membership enters the calibration process, reducing the chance of bias in the estimates. The median representative agent of each cluster is then obtained by taking the median of the parameters per cluster.

Table 4 shows the parameter estimates for the representative agents across clusters as well as for the whole sample. It is important to note that because the median value of each parameter is shown, the constraints might not hold for all rows in the table.

Figure 16 shows the estimated lie (grey) and truth (white) regions for the estimated representative individuals for the clusters defined in Section 4 above.[8]

Never-liars' estimates are characterised by taking a large value for c_θ and other parameter

[8] As an alternative approach to the median representative agents one can calculate the median behaviour of each cluster, thereby obtaining behaviourally representative agents. The model is then calibrated to each of these six behaviourally representative agents. While the resulting estimates are more representative of the clusters, the approach suffers from a sequential testing problem where, if the cluster allocation itself were flawed, the results of the calibration would be as well. For completeness, I provide the results of this alternative methodology in Section A.6 in the Appendix. The results are very similar to each other.

Table 4: Parameter values by representative agent for the full sample and by group

Type, θ	α_θ	β_θ	δ_θ	γ_θ	c_θ
Full sample	1	0	0.51	0.13	2.11
	(0.48)	(0.48)	(0.61)	(0.34)	(1.66)
Never-liar	1	0	1	0	22.20
	(0.04)	(0)	(1)	(0)	(18.62)
Max. DM's payoff	0.86	0.14	0.76	0.14	0.64
	(0.21)	(0.21)	(0.20)	(0.20)	(1.30)
Max. both payoffs	0.52	0.48	0.52	0.48	0
	(0.07)	(0.07)	(0.08)	(0.07)	(0.09)
Max. charity's payoff	0.40	0.61	0.05	0.65	0.81
	(0.41)	(0.72)	(0.29)	(0.50)	(0.97)
Balancing behaviour	0.88	0	0	0.26	1.63
	(0.55)	(0)	(0)	(0.12)	(0.86)
Non-strategic behaviour	1.11	0	0	0.15	1.72
	(0.59)	(0)	(0)	(0.32)	(1.00)

The median parameter values calculated by cluster are given for the whole sample and by cluster. The interquartile range for each parameter is shown in brackets under the median.

estimates are either equal to zero or small relative to the constant cost. The combination of the parameters ensures that these subjects do not lie in the payoff space of the experiment.

For those subjects that lie mainly when the DM gains, subjects are characterised by relatively small parameter estimates that, in combination, put the most weight on the DM's monetary payoff.

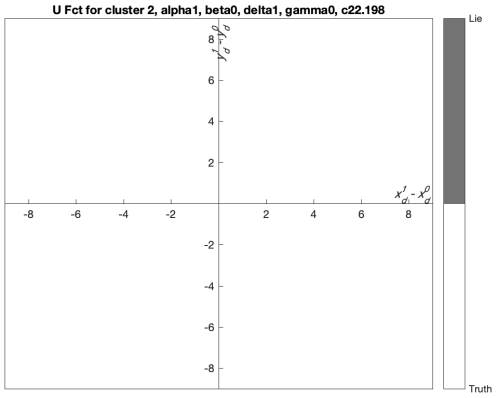
For those DMs who lie for their own as well as for the charity's gain parameter estimates are also small. However, the constant cost is close to zero and the weights on $x_d^m - x_d^0$ and $y_d^m - y_d^0$ are nearly equal for both functions of the max.

For those who lie mainly to benefit the charity, parameter estimates are similar in size to those for the previous cluster. However, as expected, the weights on the charity's payoff are larger than those on the DM's. Perhaps surprisingly, the weight on the DM's payoff is very small for the second function in the max operator. This implies that there is a kink in the lie region, which is shown in Figure 16(d).

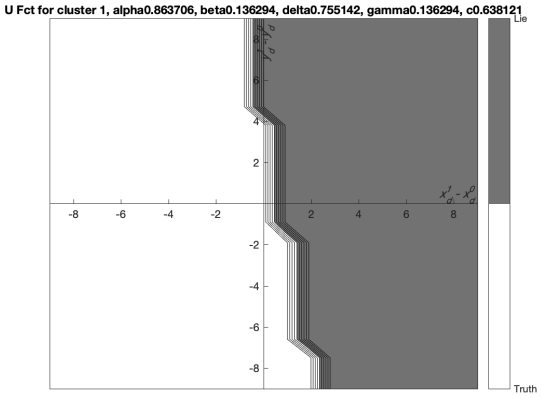
For the two groups where behaviour looks non-systematic, the suspected balancing and

Figure 16: Estimated lie regions for representative agents of groups identified in Section 4

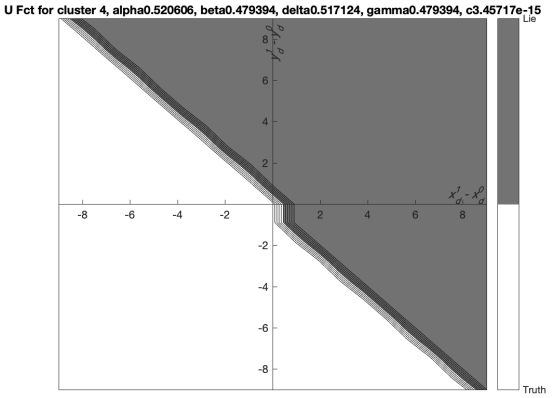
(a) Never-liar



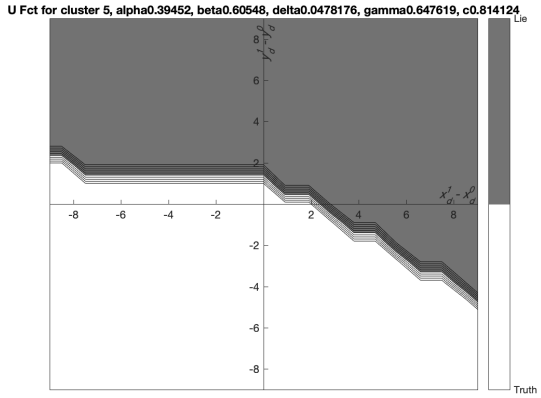
(b) Lies mainly when DM gains



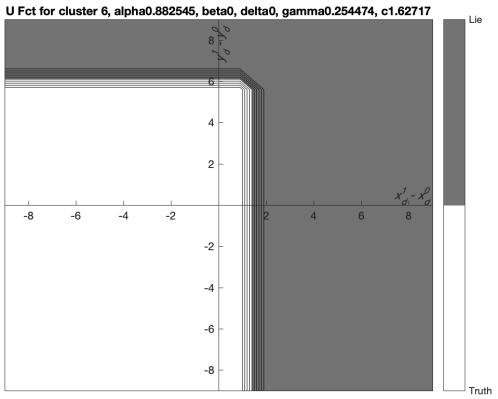
(c) Lies for both DM's and charity's gain



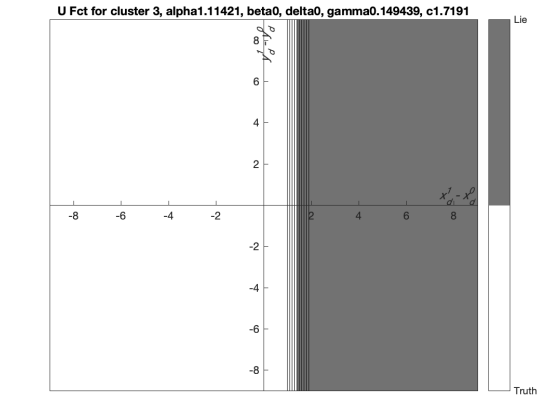
(d) Lies when the charity gains



(e) Balancing?



(f) Non-strategic behaviour



The x-axis displays $x_d^1 - x_d^0$ and the y-axis displays $y_d^1 - y_d^0$. The upper right quadrant thus displays *mutually beneficial lies*, the lower right quadrant displays *egoistic lies*, the lower left shows the *mutually harmful lies* and the upper left shows the *altruistic lies*. Weak types, for example *weakly altruistic lies*, are displayed on the axes. Dark grey areas indicate that the agent is expected to lie and white areas indicate payoff combinations for which the agent is expected to tell the truth.

the non-systematic group, the estimation procedure indicates that Properties 3(1) and (2) hold and that there is thus a kink and that behaviour is described by Property 3(1), respectively.

These lie regions provide a simple tool for the analysis and prediction of lying behaviour in binary choice settings. Based on the estimation results, such a figure can be created for every subject in the dataset. Due to the single-crossing property, Property 3, imposed by the framework it is possible to visualise how a subject is expected to behave for alternatives that were not included in the experiment. The lie regions thus provide the researcher with a visual tool to help anticipate behaviour without requiring additional data.

7. Conclusion

This paper proposes a unifying framework in which lying preferences can be analysed. The framework defines the setting and the space of lie types and provides properties of behaviour. It also informs the design of an experiment aimed at eliciting lying preferences in this setting. The results show that the identified lie types matter and that the properties impose plausible constraints on behaviour. Importantly, accounting for both improves predictive power relative to existing benchmark models. The experimental component introduces an experimental design that models lie types in a straightforward manner, permits the researcher to observe lying choices at the individual level and allows the clean separation of lying and social preferences. The design can easily be paired with modern machine learning methods which are useful for the analysis of decision-maker types. Employing a combination of principal component analysis and a k-means algorithm, I find that there are six major behavioural types of decision-makers. Knowing to which group a decision-maker belongs vastly improves the performance of out of sample predictions. In contrast, knowing how the subject behaved in a social preference game analogous to the lying game does not help to predict lying decisions. Finally, I propose a parametric model and calibrate it to describe behaviour. The need for a unifying framework of lying is underlined by the large improvements in predictive accuracy when accounting for the heterogeneity of lies and decision-makers that are the key elements of the framework.

To conclude, in this paper, I have examined the fundamentals of lying preferences in the presence of heterogeneity of lie types and decision-maker types. The resulting insights can be used for the purpose of model building and prediction analyses. Having identified and modelled the fundamental preferences, the next step is to establish their relationship with other aspects that enter the decision to lie such as time pressure concerns or reputation and probability of detection which are important in repeated interactions.

References

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

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

Bagnall, A., G. Janacek, and M. Zhang (2003): "Clustering time series from mixture polynomial models with discretised data," .

Biziou-van Pol, L., J. Haenen, A. Novaro, A. Occhipinti Liberman, and V. Capraro (2015): "Does telling white lies signal pro-social preferences?" *Judgment and Decision Making*, 10, 538–548.

Cameron, S. V. and J. J. Heckman (1998): "Life cycle schooling and dynamic selection bias: Models and evidence for five cohorts of American males," *Journal of Political economy*, 106, 262–333.

Cappelen, A. W., E. Ø. Sørensen, and B. Tungodden (2013): "When do we lie?" *Journal of Economic Behavior & Organization*, 93, 258–265.

Capraro, V. (2017): "Does the truth come naturally? Time pressure increases honesty in one-shot deception games," *Economics Letters*, 158, 54–57.

Chamberlain, G. (1983): "Funds, factors, and diversification in arbitrage pricing models," *Econometrica*, 1305–1323.

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

Engle, R. and M. Watson (1981): "A one-factor multivariate time series model of metropolitan wage rates," *Journal of the American Statistical Association*, 76, 774–781.

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

Falat, L. and L. Pancikova (2015): "Quantitative modelling in economics with advanced artificial neural networks," *Procedia economics and finance*, 34, 194–201.

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

Focardi, S. M. and T. Intertek Group (2001-2004): "Clustering economic and financial time series: Exploring the existence of stable correlation conditions," *Discussion Paper*.

Frederick, S. (2005): "Cognitive reflection and decision making," *Journal of Economic perspectives*, 19, 25–42.

Gibson, R., C. Tanner, and A. F. Wagner (2013): "Preferences for truthfulness: Heterogeneity among and within individuals," *The American Economic Review*, 103, 532–548.

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

Gneezy, U., B. Rockenbach, and M. Serra-Garcia (2013): "Measuring lying aversion," *Journal of Economic Behavior & Organization*, 93, 293–300.

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media.

Hurkens, S. and N. Kartik (2009): "Would I lie to you? On social preferences and lying aversion," *Experimental Economics*, 12, 180–192.

Kartik, N. (2009): "Strategic communication with lying costs," *The Review of Economic Studies*, 76, 1359–1395.

Kerschbamer, R., D. Neururer, and A. Gruber (2019): "Do altruists lie less?" *Journal of Economic Behavior & Organization*, 157, 560–579.

Levine, E. E. and M. E. Schweitzer (2014): "Are liars ethical? On the tension between benevolence and honesty," *Journal of Experimental Social Psychology*, 53, 107–117.

Lohse, T., S. A. Simon, and K. A. Konrad (2018): "Deception under time pressure: Conscious decision or a problem of awareness?" *Journal of Economic Behavior & Organization*, 146, 31–42.

Palan, S. and C. Schitter (2018): "Prolific. ac?A subject pool for online experiments," *Journal of Behavioral and Experimental Finance*, 17, 22–27.

Peer, E., L. Brandimarte, S. Samat, and A. Acquisti (2017): "Beyond the Turk: Alternative platforms for crowdsourcing behavioral research," *Journal of Experimental Social Psychology*, 70, 153–163.

Rammstedt, B. and O. P. John (2007): "Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German," *Journal of research in Personality*, 41, 203–212.

Stock, J. H. and M. W. Watson (2005): "Implications of dynamic factor models for VAR analysis," Tech. rep., National Bureau of Economic Research.

Thomson, K. S. and D. M. Oppenheimer (2016): "Investigating an alternate form of the cognitive reflection test," *Judgment and Decision making*, 11, 99.

A. Appendix

A.1. Instructions of the Experiment

A.1.1. Introduction

Thank you for deciding to participate. The experiment is split into two large blocks of questions. Each block contains 60 short rounds. Your choices in the experiment will affect your bonus payments and that of a partner. Specifically, this partner is a charity that you can select from the list below. At the end of the experiment, one question from each block will be selected for payment purposes by chance. You will receive the payment as a bonus payment via Prolific and the charity will receive the payment through a donation. Payoffs will be displayed in terms of tokens. These tokens will be converted into GBP at the end of the experiment. 5 tokens correspond to GBP1. Before you start each block, you will receive detailed instructions for that block.

Which charity would you like to be partnered with? Please select one.

Please select a charity before you continue.

- Cancer Research UK
- Children in Need
- Comic Relief
- National Trust
- WWF

A.1.2. Instructions - Lying game

Please read the instructions carefully.

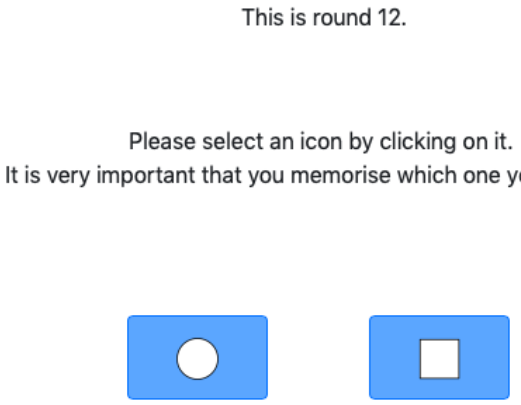
In this block, you will play 60 rounds of the following format: Each round consists of 2 screens, displayed directly after one another.

First screen:
You have to select one icon by clicking on the button on which the icon is displayed.
There exist 6 icons, each a geometrical shape, in total: square, circle, triangle, diamond, pentagon and hexagon.

It is very important that you memorise which icon you have selected.

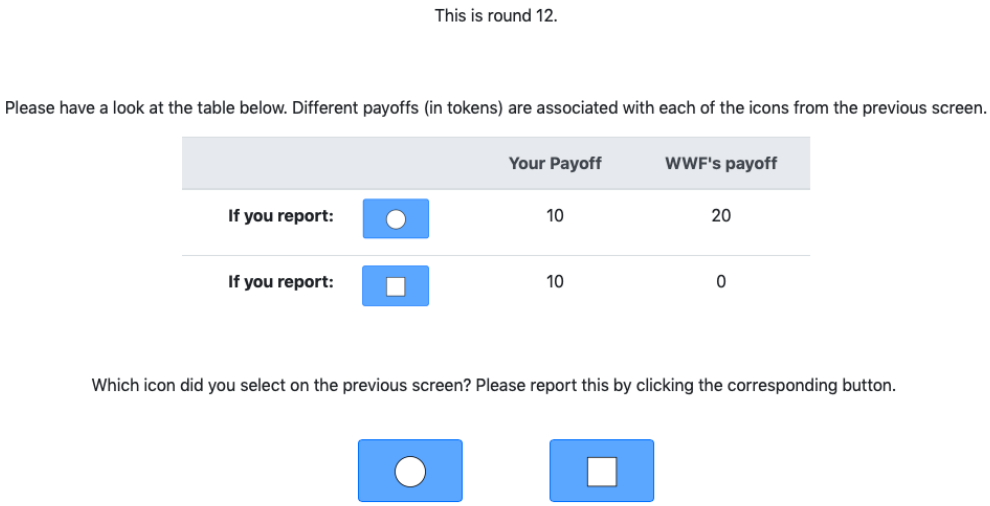
Below, you can see an example of the first screen:

Figure 17: Example screen of the 1st screen of one of the rounds of the lying game.



Second screen:
After having selected the icon on the previous screen, you will see a screen similar to the one below:

Figure 18: Example screen of the 2nd screen of one of the rounds of the lying game.



To each icon from the first screen correspond 2 token values.
In the table, you can see the token values of each icon by looking at the row that shows

the respective icon.
The first column shows your monetary value and the second column the charity's monetary value.
Importantly, the icons that are displayed on the buttons are drawn randomly each round which means that the token values associated with a particular icon vary each round.

Decision:
At the bottom of the second screen, you have to report which icon you selected in the first screen.
Your payoff from this round will be determined by the icon you report on the second screen and not the one you selected on the first screen.
We understand that each round, you may have reasons for or against reporting truthfully.

Payments:
In this block, one round will be selected by chance. You and the charity that you selected will receive the payments according to the button that you reported in this round.

Please answer the following questions to confirm that you have understood the instructions.

You have to answer all questions correctly before you can continue with the experiment.

1) Imagine that you selected the square on the first screen (in the image above). Then, on the second screen (in the second image above), you clicked on the circle. Which of the following statements is true?

- I told the truth when I reported the circle and the charity will receive a larger payment than if I had clicked on the square.
- I told the truth when I reported the circle and the charity will receive a smaller payment than if I had clicked on the square.
- I told a lie when I reported the circle and the charity will receive a larger payment than if I had clicked on the square.
- I told a lie when I reported the circle and the charity will receive a smaller payment than if I had clicked on the square.

2) With whom will you be matched during the experiment?

- A charity of my own choosing
- A computer generated player

3) What will your payment depend on?

- The icon that you selected on the first screen
- The icon that you reported on the second screen

A.1.3. Instructions - Social preference game

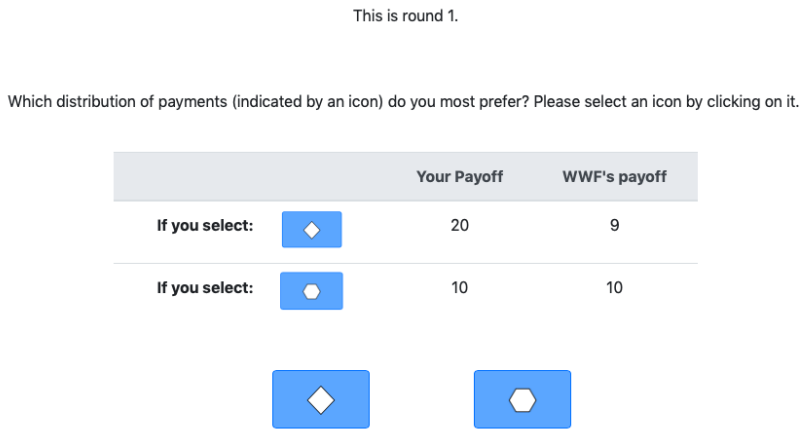
Please read the instructions carefully.

In this block, you will play 60 rounds where each round consists of exactly one question screen.

Question screen:
You will see a table where each row corresponds to a choice object and each column to monetary payoffs (in tokens).

The first column shows your monetary gain from selecting that icon and the second column the charity's monetary gain. Please take a look at the example screen below:

Figure 19: Example screen of one of the rounds of the social preference game.



Decision:
At the bottom of the screen, you have to select which of the options i.e. icons you prefer. Your payoff and that of the charity from this round will be determined by the icon that

you click on.
There exist 6 icons, each a geometrical shape, in total: square, circle, triangle, diamond, pentagon and hexagon.
The symbols are allocated randomly to a payoff each question. For example, that means that the payoff pair "10 for you and 5 for the charity" could have any of the icons next to it.

Payments:
In this block, one round will be selected by chance. You and the charity that you selected will receive the payments according to the icon that you selected in this round.

Please answer the following questions to confirm that you have understood the instructions.

You have to answer all questions correctly before you can continue with the experiment.
1) With whom will you be matched during the experiment?

- A charity of my own choosing
- A computer generated player

2) In the example above, which choice will maximise your payoff?

- Clicking on the button displaying the diamond
- Clicking on the button displaying the hexagon

3) In the example above, which choice will maximise the charity's payoff?

- Clicking on the button displaying the diamond
- Clicking on the button displaying the hexagon

A.2. Specification of rounds of the experiment

Table 5: Rounds with non-binary choice

Types of lies	(x^0, y^0)	(x^1, y^1)	(x^2, y^2)	(x^3, y^3)	(x^4, y^4)	(x^5, y^5)
MHL	(5, 6)	(4, 4)				
HL	(5, 6)	(5, 5)				
SHL	(11, 10)	(10, 10)				
WAL	(10, 10)	(10, 13)				
WAL	(10, 8)	(10, 13)				
WAL	(10, 9)	(10, 10)				
WAL	(10, 11)	(10, 13)				
SSL	(10,10)	(11, 10)				
SSL	(10,10)	(15,10)				
SSL	(10, 9)	(12, 9)				
SSL	(7, 10)	(10, 10)				
MBL	(14,12)	(15, 15)				
MBL	(10, 8)	(14, 10)				
MBL	(10, 8)	(11, 9)				
MBL	(8, 10)	(11, 11)				
MBL	(7, 9)	(12, 10)				
MBL	(7, 10)	(9, 12)				
MBL	(8, 10)	(10, 13)				
MBL	(10, 9)	(12, 13)				
MBL	(9, 7)	(10, 12)				
MBL	(6, 10)	(10, 14)				
MBL	(7, 7)	(10, 10)				
EL	(10, 10)	(12, 8)				
EL	(10, 8)	(11, 7)				
EL	(10, 15)	(11, 10)				
EL	(10, 15)	(15, 10)				
EL	(9, 12)	(10, 9)				
EL	(9, 10)	(12, 9)				
EL	(9, 12)	(11, 11)				
EL	(10, 10)	(13, 8)				
EL	(7, 13)	(10, 10)				
EL	(7, 11)	(12, 10)				

EL	(6, 14)	(9, 9)				
EL	(11, 10)	(15, 7)				
EL	(2, 10)	(7, 8)				
EL	(5, 10)	(13, 9)				
AL	(10, 10)	(9, 11)				
AL	(10, 5)	(9, 8)				
AL	(10, 9)	(9, 15)				
AL	(10, 6)	(8, 8)				
AL	(15, 5)	(13, 13)				
AL	(10, 10)	(8, 14)				
AL	(14, 8)	(12, 13)				
AL	(10, 10)	(7, 13)				
AL	(12, 4)	(10, 10)				
AL	(14, 6)	(10, 10)				
AL	(15, 5)	(12, 11)				
AL	(14, 7)	(11, 12)				
AL	(15, 4)	(12, 12)				
AL	(16, 5)	(14, 12)				
MBL, EL	(10, 10)	(15, 15)	(18, 9)			
MBL, EL	(10, 10)	(11, 11)	(13, 9)			
AL, EL	(10, 10)	(9, 15)	(15, 9)			
EL, AL	(12, 10)	(15, 9)	(11, 15)			
SSL, WAL	(10, 10)	(15, 10)	(10, 15)			
SSL, WAL	(5, 10)	(10, 10)	(5, 11)			
WAL, AL	(10, 10)	(10, 12)	(9, 15)			
SSL, EL	(10,10)	(12, 10)	(15, 9)			
MBL, SSL, WAL, AL, EL	(10, 10)	(11, 11)	(12, 10)	(10, 12)	(9, 15)	(15, 9)
MBL, SSL, WAL, AL, EL	(8, 10)	(11, 11)	(8, 15)	(6, 18)	(13, 10)	(16, 8)

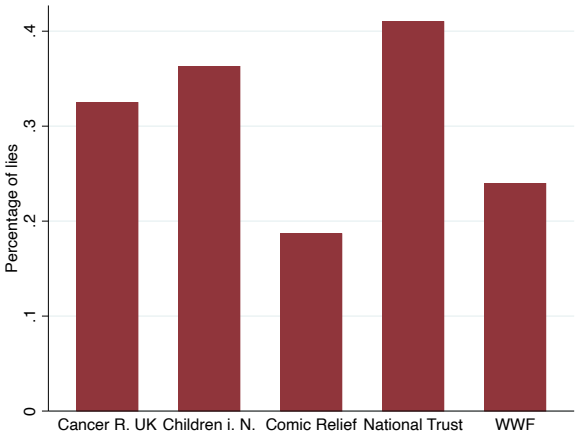
The table displays the payoff bundles for all experimental rounds with more than two states. (x^0, y^0) refers to the payoffs of the true state; all other payoff bundles are linked to untrue states that are available for reporting in addition the true state. The acronyms stand for *mutually beneficial lie* (MBL), *weakly altruistic lie* (WAL), *altruistic lie* (AL), *egoistic lie* (EL), *self-serving lie* (SSL), *self-harming lie* (SHL), *harmful lie* (HL) and *mutually harmful lie* (MHL).

A.3. Details on comparison of aggregate results to the literature

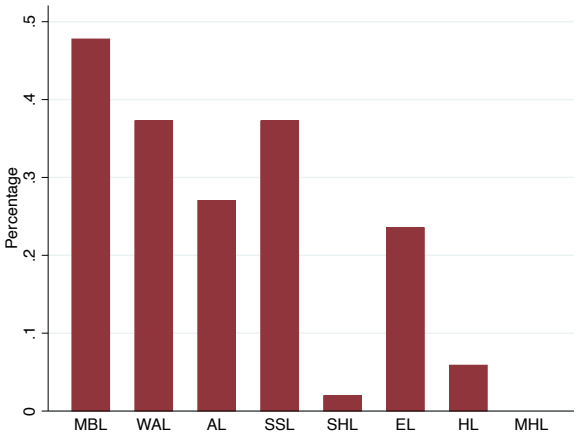
A.3.1. Lying behaviour across charities

Figure 20: Average percentage of lies by lie type for each charity

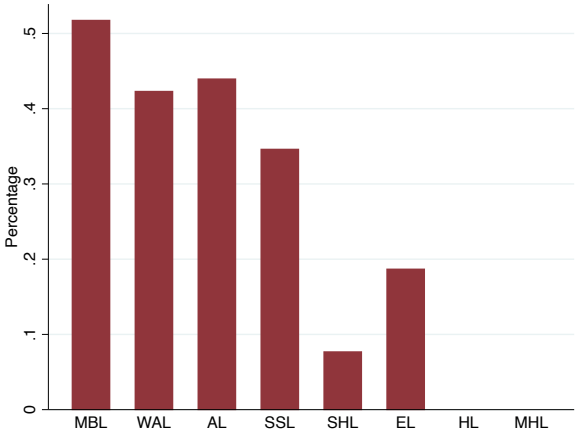
(a) Average number of lies by charity



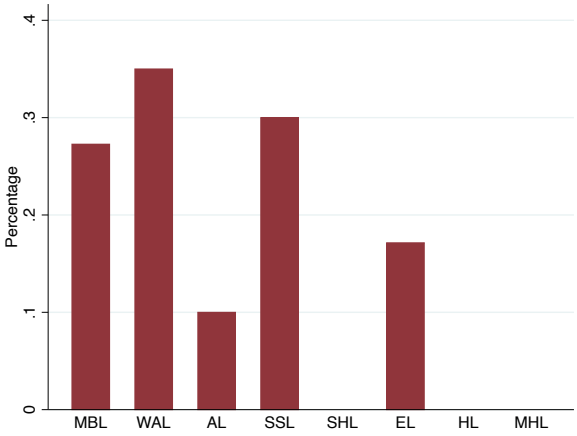
(b) Cancer Research UK



(c) Children in Need

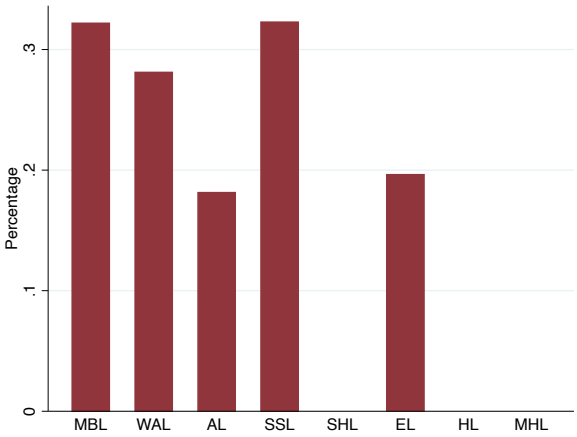


(d) Comic Relief



(e) National Trust

(f) World Wildlife Fund



Panel (a) shows the average percentage of lies told by subjects who selected the charity. Each bar represents one charity. Panels (b) - (f) show the percentage of lies by lie type for for each of the charities. The acronyms stand for *mutually beneficial lie* (MBL), *weakly altruistic lie* (WAL), *altruistic lie* (AL), *self-serving lie* (SSL), *self-harming lie* (SHL), *egoistic lie* (EL), *harmful lie* (HL) and *mutually harmful lie* (MHL).

A.3.2. Covariates of lying

After completion of the two main stages of the experiment, subjects were asked to answer a series of questions and to perform cognitive ability as well as personality tests. The following paragraphs examine whether there exists a relationship between certain personality or cognitive traits as well as demographics and lying behaviour. Results are displayed in Table 6.

A commonly discussed feature in the literature is gender. I find a significant difference in the average percentage of lies between men and women (t-test, p-value<0.01) overall as well as for most of the lie types. Specifically, I find that men lied significantly more (36.33% compared to 26.23% across all questions). The same relationship exists for *mutually beneficial lies (MBLs)*, *weakly altruistic lies (WALs)*, *self-serving lies (SSLs)* and *egoistic lies (ELs)*. Not surprisingly, there is no difference for *self-harming lies (SHLs)*, *harmful lies (HLs)* and *mutually harmful lies (MHLs)* as nearly no subject lied for these. Interestingly, there exists no significant difference in the behaviour for *altruistic lies (ALs)*. The finding that men lie more in MBLs is in line with Erat and Gneezy (2012) but goes against the finding of no differences in Biziou-van Pol et al. (2015) and Cappelen et al. (2013). The finding that more men lie for ELs is in line with Dreber and Johannesson (2008) while the finding of no differences for ALs goes against Erat and Gneezy (2012)'s finding that more women tell ALs. Yet, it is noteworthy that ALs are the only type of lies for which women lied more than men (except for HLs and SHLs which are likely to be due to errors) which, while statistically insignificant, does provide some evidence in the direction of Erat and Gneezy (2012), especially when comparing this to the large differences between men and women's percentage to lie for the other lie types. Importantly, previous papers often asked a very low number (sometimes only one) of lie questions so that it is perhaps not surprising that those results are volatile when using different question specifications.

A second feature of interest is cognitive ability. Subjects answered a CRT test (see Thomson and Oppenheimer (2016) and Frederick (2005)) where they received one point for each correct answer out of a total of four questions. Testing lie behaviour for each level of the score as well as for measures that split the sample into two groups based on their score (two different splits were used) reveals that subjects with different scores answered questions significantly differently both on average and for each of the lie types. Interestingly, subjects with a higher CRT score lied significantly more for each lie type except for HLs where subjects with lower scores lied more (this is likely to be caused by choice errors of which subjects with low CRT scores seem to perform more).

Subjects also answered a ten item Big Five personality test (Rammstedt and John (2007)). Splitting subjects into two groups (high versus low performing) for each of the big five character traits, I analyse if there exist systematic differences. For example, subjects who scored more than five out of ten points on the conscientiousness measure lied significantly less than people who scored five or fewer points. This also holds when looking at MBLs, WALs, SSLs and ELs, specifically. However, there is no difference in behaviour for ALs. Conscientiousness could be linked to being more rule-abiding. This could then explain the lower number of lies even when these lies help the charity.

Table 6: Significant differences in group behaviour by covariates of interest

Covariates	Average	MBLs	WALs	ALs	SSLs	SHLs	ELs	HLs
Gender	***	***	***		***		***	
Male	36.33%	54%	42.5%	24.14%	46.5%	0%	30.43%	2%
Female	26.23%	36.36%	28.3%	24.66%	28.77%	3.78%	18.87%	3.78%
CRT	***	***	***	***	**		***	
2 or fewer points	25.11%	35.40%	27.13%	18.54%	30.32%	2.13%	21.43%	6.38%
3 or 4 points	36.19%	52.92%	41.96%	29.34%	43.30%	1.79%	27.04%	0%
Conscientiousness	***	***	***		***		***	
5 or fewer points	40.40%	57.14%	42.86%	25.51%	54.76%	4.76%	38.10%	0%
6 to 10 points	28.76%	41.80%	33.23%	24.13%	32.93%	1.22%	20.99%	3.66%

Group behaviour is defined as the percentage of lies for each lie type of a group. Stars indicate significance levels with p<0.1 *, p<0.05**, p<0.01***. The acronyms stand for *mutually beneficial lie* (MBL), *weakly altruistic lie* (WAL), *altruistic lie* (AL), *self-serving lie* (SSL), *self-harming lie* (SHL), *egoistic lie* (EL) and *harmful lie* (HL). *Mutually harmful lies* are not displayed as the percentage of lies was 0% for all comparisons.

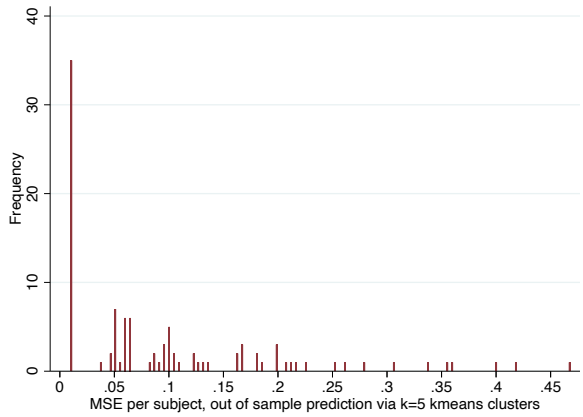
A.4. Supporting results for Section 4

A.4.1. Robustness to misspecification of number of clusters k

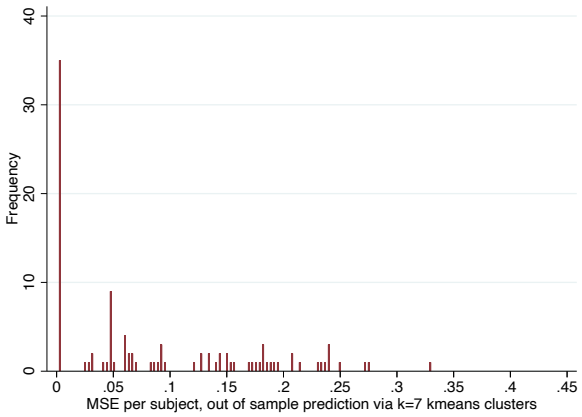
The figure shows the results from rerunning the heterogeneity out of sample forecasting exercise with $k = 5$ and $k = 7$. The comparison to Figure 11(b) shows that the gains from accounting for heterogeneity in lying preferences are robust to slight misspecification of k.

Figure 21: Histogram of each subject's MSE in pseudo out of sample prediction exercise with heterogeneity in preferences taken into account for different k.

(a) $k = 5$ clusters



(b) $k = 7$ clusters



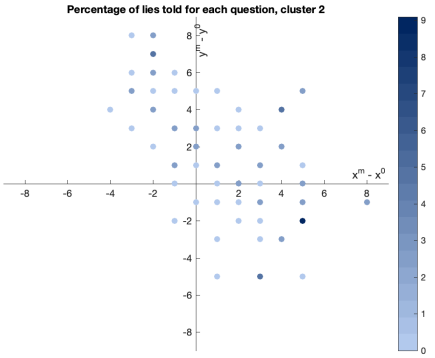
A.4.2. Never-liars and potential errors of behaviour

The k-means algorithm clustered 44 subjects into the never-liar group. Of these, 28 never misreported. The remaining subjects misreported one to 7 times. While I cannot rule out alternative explanations, these misreports appear to be honest mistakes rather than lies. First, shifts appear to be random rather than systematic even examining alternative explanations such as inequality concerns. Second, the k-means algorithm identified them as never-liars.

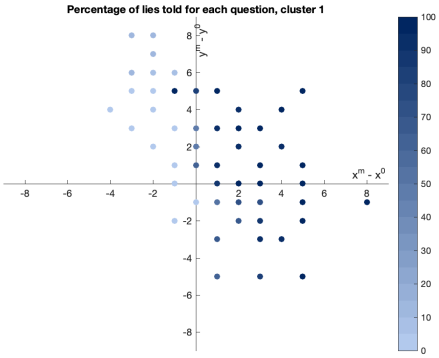
A.4.3. Behaviour of clusters

Figure 22: Percentage of lies told for each question in the lying game by cluster

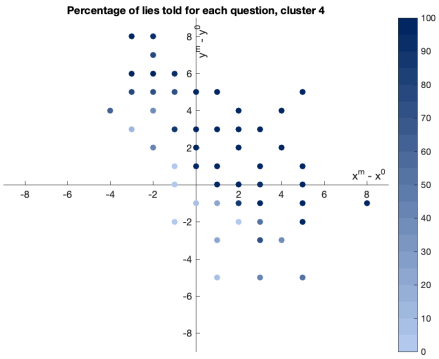
(a) Never-liar



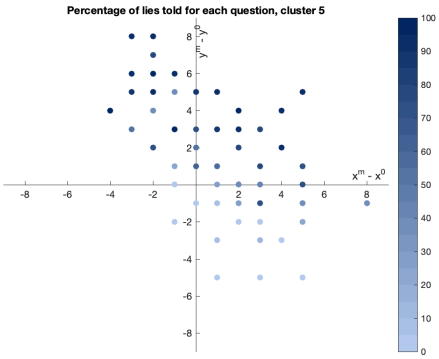
(b) Lies mainly when DM gains



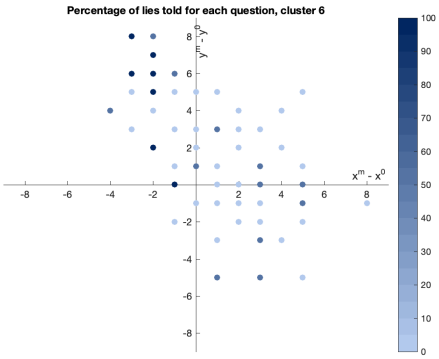
(c) Lies for both DM's and charity's gain



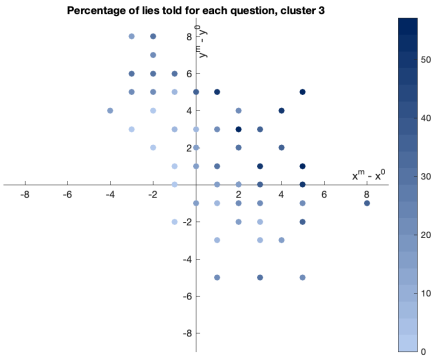
(d) Lies when the charity gains



(e) Balancing?



(f) Non-systematic behaviour

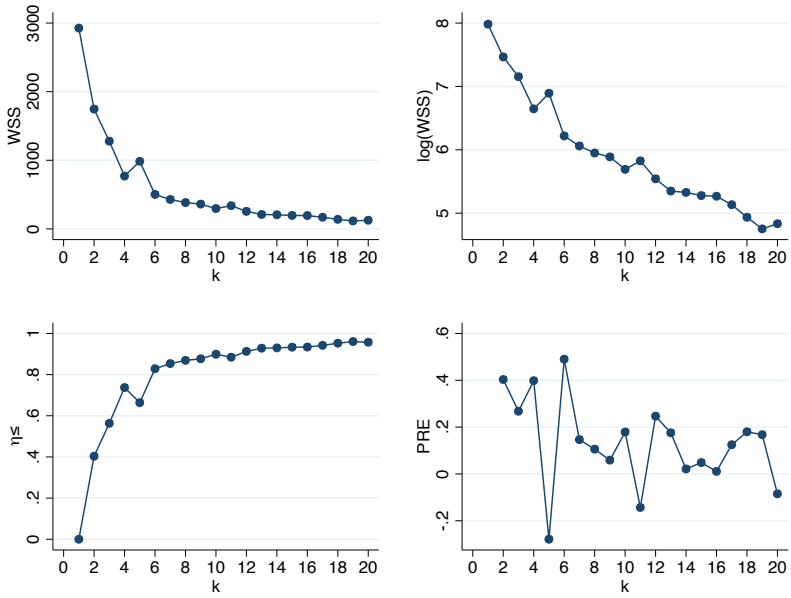


The x-axis displays $x^1 - x^0$ and the y-axis displays $y^1 - y^0$. Each dot represents a round with a binary choice of the lying game expressed via the payoff combinations. Lighter blues indicate a higher percentage of truths told and darker blues a higher percentage of lies told. The color scale to the right links the percentage to the color.

A.5. Supporting results for Section 5

For robustness of the k-means analysis, I conducted a pre-analysis to identify the ideal number of k. The results here are not as clear as in the lie k-means analysis as explanatory power increases when having more than six groups. However, even with six groups, there are two groups that contain between one and three subjects which indicates that the clustering is already quite detailed. For that reason, it seems that six groups are satisfactory. Figure 23 shows summary statistics for the performance of 1 to 20 clusters.

Figure 23: Performance indicators by cluster number in social preference game k-means analysis



Plot of weak sum of squares (WSS), log WSS , η^2 and proportional reduction of error (PRE) for number of clusters $k = \{1, ..., 20\}$.

To rule out that misspecification of the number of clusters is driving the low NMI score, I reran the analysis with $k = 4$ clusters.[9] Specifically, the k-means analyses from Sections 4 and 5 were repeated with k=4 and the NMI score was calculated for the new clusters. The resulting NMI score was equal to 0.1998, which is very similar to the NMI score obtained from comparing the clusters of the k-means analyses with $k = 6$. This suggests that the finding that cluster membership does not overlap between the lying and the social preference games is robust to the number of clusters specified and thus increases confidence in this finding.

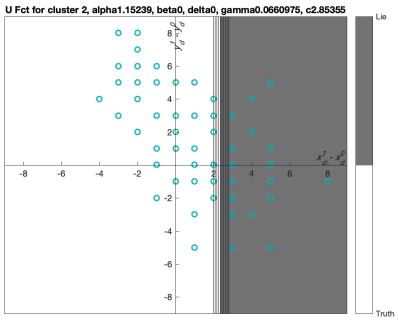
[9]The fact that two groups have only between one and three members makes $k = 4$ the most likely contender.

A.6. Supporting results for Section 6

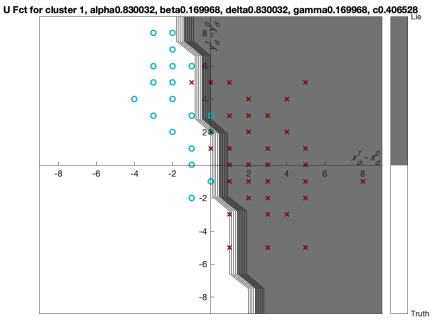
Alternative calibration exercise.

Figure 24: Estimated lie regions for representative agents based on the mean behaviour of each cluster identified in Section 4

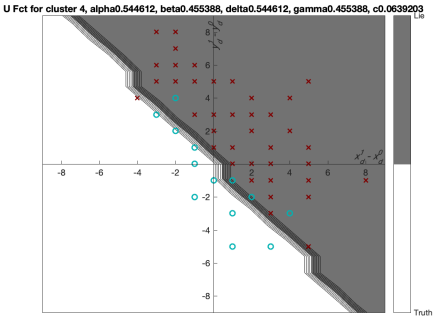
(a) Never-liar



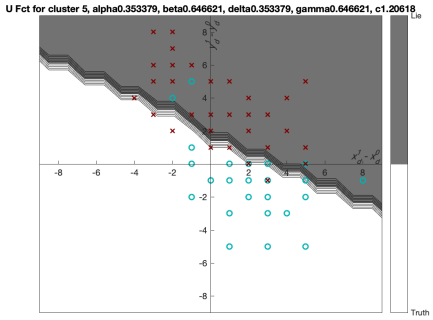
(b) Lies mainly when DM gains



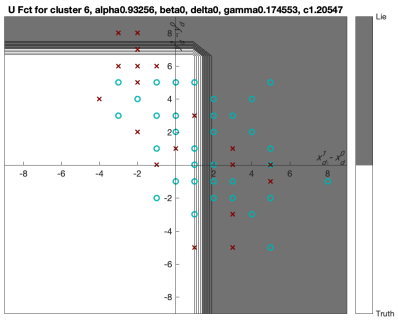
(c) Lies for both DM's and charity's gain



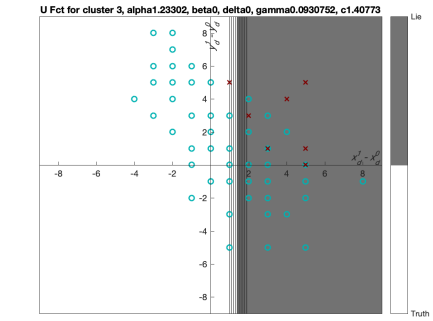
(d) Lies when the charity gains



(e) Balancing?



(f) Non-systematic behaviour



The x-axis displays $x^1 - x^0$ and the y-axis displays $y^1 - y^0$. The upper right quadrant thus displays *mutually beneficial lies*, the lower right quadrant displays *egoistic lies*, the lower left shows the *mutually harmful lies* and the upper left shows the *altruistic lies*. Weak types, for example *weakly altruistic lies*, are displayed on the axes. Dark grey areas indicate that the agent is expected to lie and white areas indicate payoff combinations for which the agent is expected to tell the truth.

A.7. Glossary for machine learning methodology

This section provides definitions of the terminology used in Sections 4 and 5 to describe the machine learning (ML) methodology used in the paper.

- *Label:* The dependent variable.
- *Unsupervised learning:* A ML algorithm that transforms data (for example predicts or groups data) where the true labels i.e. the value of the dependent variable are unknown to the researcher. One can imagine the difference between supervised and unsupervised learning to be parallel to within versus out of sample predictions. In sample, the value of the dependent variable is known so that the researcher can compare the performance of the prediction against the true values; this is similar to supervised learning. Out of sample, the value of the dependent variable is forecasted but is unknown to the researcher; this is akin to unsupervised learning.
- *Cluster:* A group of observations that are most similar to each other.
- *Centroid:* A centroid is the center of a cluster. This location does not have to be the real center as it is often initially allocated randomly.
- *k-means clustering:* An unsupervised ML algorithm where data is sorted into k clusters. The algorithm is initially allocated random centroids for each cluster. Based on the distance of a set of pre-specified independent variables (also called features in the ML literature) to the centroid, observations are allocated to each cluster. The centroids are then updated in that they are moved to the mid position of the features of the observations that had been allocated to the cluster in the previous step. The observations are then reallocated based on which of the updated centroids is the closest to them. This procedure is repeated until updating the centroids does not lead to a change in allocations any more.
- *Principal component/factor:* A weighted, often linear, combination of correlated independent variables.
- *Principal components analysis (PCA):* The (unsupervised) process of identifying the principal components/ factors in the data. As the components are ordered by how much of the variance in the data they can explain, they can be used to reduce the number of variables needed in an analysis.
- *Factor model analysis:* The process of identifying the principal components/ factors that describe the most variance in the data and restricting the model to these factors.