A survey of belief-based guilt aversion in trust and dictator games

Edward Cartwright[*†‡]

14th April 2018

Keywords: Guilt aversion, social norms, trust game, belief, dictator game.

JEL codes: C72, D03, C92

Abstract

The evidence for belief-based guilt aversion is reviewed with a particular focus on trust games and dictator games. By way of comparison an alternative model to belief-based guilt aversion is proposed which is based on an internalized norm. We show that the experimental evidence to date is consistent with belief-based guilt aversion but that it is difficult to distinguish one model from another. The review compares the many different approaches that have been used to elicit beliefs. It also looks at the role of exposure and communication.

[*] Department of Strategic Management and Marketing, Leicester Castle Business School, De Montfort University, Leicester, LE1 9BH, UK. Corresponding author, email edward.cartwright@dmu.ac.uk.

[†] I would like to thank Martin Dufwenberg (co-editor) and two anonymous referees for constructive and very useful comments on an earlier version of the paper. I would also like to thank the audience at the 2nd Workshop on Behavioral Game Theory for their feedback, and the organizers for putting on a great Workshop.

[‡] The final version of this paper is to be published in the *Journal of Economic Behavior and Organization*, Elsevier.

1 Introduction

The basic idea behind belief-based guilt aversion is that a person feels guilt if they believe they have let down the payoff expectations of another. For instance, someone who does not fulfill a contract may feel guilt if they believe the other party to the contract expected the contract would be fulfilled. A general model of belief-based guilt aversion was formally introduced by Battigalli and Dufwenberg (2007), drawing on prior work developed for a range of specific contexts (e.g. Huang and Wu 1994, Dufwenberg and Gneezy 2000, Dufwenberg 2002, Guerra and Zizzo 2004 and Charness and Dufwenberg 2006).[1] There have now been a large number of papers providing experimental evidence for and against the belief-based model's predictions (e.g. Vanberg 2008, Ellingsen et al. 2010, Kawagoe and Narita 2014). So what, if anything, have we learnt so far? That's the question I attempt to answer in this paper.

In order to review the evidence on guilt aversion I first propose a model of guilt that can be readily compared with the belief-based model. To put this contribution in context it is important to recognize that the belief-based model of guilt aversion, proposed by Battigalli and Dufwenberg (2007), is widely applicable. For instance, it can used to study guilt from free-riding on the production of a public good (Dufwenberg, Gächter and Hennig-Schmidt 2011), guilt from over-charging for a credence good (Beck et al. 2013) or guilt from lying (Battigalli, Charness and Dufwenberg 2013). It is no surprise that a model as widely applicable as this does not always perfectly predict what we observe in the lab. The interesting question is whether some other model can capture aspects of guilt that the belief-based model cannot.

The alternative model that I shall consider is similar to that discussed by Lopez-Perez (2010) and sees guilt as arising from deviation from some internalized norm. In narrowing down what norm means I will follow the approach of Andrighetto, Grieco and Tummolini (2015) in distinguishing between empirical and normative expectations. In a model based on *empirical conformity* a person's guilt is relative to what they believe the 'average person' in their situation would do. For instance, someone feels guilt from not fulfilling a contract if they believe the 'average' person would fulfill the contract. In a model based on *normative beliefs* a person's guilt is relative to

[1] This prior work used terms like remorse (Huang and Wu 1994), letting down (Dufwnberg and Gneezy 2000) or shame (Guerra and Zizzo 2004) but the underlying ideas are similar. Dufwenberg (2002) explicitly talks of belief-dependent guilt.

what they think someone in their situation 'ought to do'. In this case the person feels guilt if they believe that fulfilling the contract is the right thing to do.

To illustrate the models and the differences between them consider the following example. Ann has employed Brian to do a job. Brian's effort, measured in time, is unobservable and so there is an incentive for him to shirk. Suppose that Brian believes the right thing to do is work for 30 minutes, that the average person would work for 20 minutes and (he believes) Ann only expects him to work for 10 minutes. If Brian works for 10 minutes then he will feel guilt according to empirical conformity and normative beliefs, because he has put in less effort than his reference level, but would not feel guilt according to the belief-based model, because he has behaved consistent with what he believes Ann was expecting. The key point of departure between the models is that in the belief-based model Brian is focused on what *Ann* expects while in the reference-based model he is focused on what the *average person* does or what he thinks is the *right thing to do*.[2]

Having briefly introduced the different models of guilt I shall look back in detail through key experimental findings in order to evaluate the evidence for and against belief-based guilt aversion. A particular focus shall be given to the trust game introduced by Charness and Dufwenberg (2006). In short, I shall argue that both the belief-based and reference-based models appear consistent with the available evidence. This is not to say that all the evidence is supportive of both models, it clearly is not; but there is no piece of evidence so compelling that it justifies ruling out one model over the other (or rejecting both). It is, though, worth highlighting that in many settings the belief-based and reference-based models give identical predictions. For instance, if Brian and Ann know nothing about each other then Brian's beliefs of what Ann expects should coincide with his beliefs about what the 'average person' does. To better distinguish between models we will need, therefore, to consider new experimental designs.[3] The concluding discussion offers some suggestions in

[2] In the example, Ann expects less than Brian's internalized norm but things can clearly go the other way. For instance, if Brian believes that Ann is expecting 50 minutes work then by working for 30 minutes he would feel guilt according to the belief-based model but not with empirical conformity or normative beliefs.

[3] The psychology and sociology literature have a lot to say on the meaning of guilt (see Lopez-Perez 2010 for a review). If we decide that guilt is an emotion a person experiences when they let down the expectations of others then one could argue, as a matter of semantics, that the belief-based model is the correct one. In economics, the

this regard.

Before we begin the analysis let me emphasize that comparing between models is not a zero-sum game. In particular, I will point to the benefits of a hybrid approach that combines elements of both the reference-based and belief-based models. The idea that behavior is driven by multiple factors is not novel (e.g. Dufwenberg et al. 2011, Battigalli, Charness and Dufwenberg 2013). But, the interaction between different factors opens up new possibilities. To illustrate the point consider again the Ann and Brian example where Brian's internalized norm is 30 minutes of work and he believes that Ann expects 50 minutes. Brian could interpret the high expectation of Ann as unreasonable and, therefore, feel no guilt from disappointing those expectations. Or, Brian could interpret the high expectation of Ann as a signal she considers him trustworthy and so feel relatively more guilt from disappointing her expectations. Importantly, in both these scenarios it is the difference between the internalized norm and Ann's expectation that matters.

The paper proceeds as follows. In Section 2 a theoretical background is provided in which models of guilt are introduced and compared. In Sections 3 and 4 a survey is provided of the experimental evidence on guilt, with a particular emphasis on trust and dictator games. Section 5 concludes.

2 Theoretical Preliminaries

To focus the discussion I will consider a simple example. Let me emphasize that my objective in doing so is to set the scene for a review of the experimental evidence and so the analysis will be rudimentary. For a more detailed look at the theory behind the belief-based model of guilt aversion see Battigalli and Dufwenberg (2007, 2009), Khalmetski, Ockenfels and Werner (2015), Attanasi, Battigalli and Manzoni (2016) and Attanasi, Battigalli and Nagel (2016). Note also that I shall focus, as is standard in the experimental literature (e.g. Charness and Dufwenberg 2006), on an individual level analysis in which we study the incentives of a particular individual tacking their beliefs as given. It is not assumed, for instance, that one person's beliefs about another be correct or that outcomes be in equilibrium.

The example is as follows: Brian (who I will label B) has been employed

focus is clearly more on accurate predictions of behavior and so we want a theory that reliable captures choice. But then the fact that alternative models tend to give the same prediction seems less of a problem.

by Ann (who I will label A) to do a job. The choice that Brian has to make is how much effort to exert. Let $e \geq 0$ denote the amount of effort that he chooses and $c(e)$ the associated cost. Assume that c is an increasing function of effort. The issue of whether Ann can observe the effort of Brian is left open at this stage. But it will be assumed that Ann pays a fixed wage that is independent of effort. Ann is, thus, powerless to act even if she does observe low effort. Suppose that Ann's expected payoff is given by $\pi(e)$ where π is an increasing function of Brian's effort.

Before we begin the analysis let me briefly relate this example to the games commonly studied in the experimental literature. Most attention has been on dictator and mini-trust games. In mapping our example to a dictator game we can think of Brian as the dictator and Ann as the receiver. Choice of effort is then the share that Brian gives to Ann. In mapping the example to a mini-trust game (or 'lost wallet game') we can think of Brian as the trustee and Ann the trustor. (In a mini-trust game the trustor makes a binary decision to either trust or not and then conditional on being trusted the trustee makes his decision. Our example, therefore, focuses on what happens if Brian is trusted by Ann.) Choice of effort is then the amount that Brian wants to 'return' to Ann.[4]

2.1 Different models of guilt

Belief-based guilt posits that Brian will feel guilty if he believes he has exerted less effort than Ann was expecting. To keep things simple I will focus here on spot beliefs. Then we can denote by λ_A the level of effort that Ann believes Brian will exert. This is her first-order belief. And denote by λ_B Brian's (spot) belief about λ_A. This is Brian's second-order belief. If $e < \lambda_B$ then Brian feels guilt. The amount of guilt he feels will be related to how much lower Ann's payoff is than expected. As an example, we could write Brian's utility function as[5]

$$u_{bb}(e, \lambda_B) = -c(e) - \gamma_{bb} \max\{\pi(\lambda_B) - \pi(e), 0\} \quad (1)$$

[4] One issue that I should mention is whether Brian's choice set is continuous or binary. In our example it is continuous, as in some experiments (e.g. Dufwenberg and Gneezy 2000). In many experiments choice is binary (e.g. Charness and Dufwenberg 2006).

[5] More generally beliefs can be modeled as a probability distribution (Battigalli and Dufwenberg 2007). Guilt still depends on the gap between realized and expected payoff.

where γ_{bb} is a measure of Brian's sensitivity to guilt. Note that Brian's payoff depends on both his action, e, and his belief, λ_B. This brings the belief-based model under the umbrella of psychological game theory (Geanakoplos, Pearce and Stacchetti 1989, Battigalli and Dufwenberg 2009).[6] Also note that the belief-based model predicts that Brian's effort will be increasing in his second-order belief, λ_B (Dufwenberg and Gneezy 2000, Charness and Dufwenberg 2006).[7]

Contrast belief-based guilt with what I will call *reference-based guilt.* The basic idea here is that Brian feels guilty if he exerts less effort than some reference level. This, of course, requires us to say what the reference level is. I will get to this shortly. For now let us just take as given a reference level δ. If $e < \delta$ then Brian feels guilt because he exerts less effort than the reference level. As an example, we could write Brian's utility function as

$$u_{sr}(e, \delta) = -c(e) - \gamma_{sr} \max \left\{ \pi\left(\delta\right) - \pi\left(e\right), 0 \right\} \tag{2}$$

where γ_{sr} again measures sensitivity to guilt. The reference-based model predicts that Brian's effort will be increasing in his reference level δ.

Recall that in setting out the model I left it open whether or not Ann can observe the effort of Brian. It is worth highlighting (more on this in Section 4.1) that Battigalli and Dufwenberg (2007) distinguish between two different models of belief-based guilt - simple guilt and guilt from blame (see also Bacharach et al. 2007). With simple guilt Brian feels guilt irrespective of whether Ann observes his effort. With guilt from blame Brian only feels guilt if Ann observes his effort. In application this is an important distinction and suggests, more generally, that observability of actions may be a key driver in guilt influenced behavior. We will discuss this issue in full in Section 4. For now I will merely point out that one can also distinguish between simple guilt and guilt from blame in modeling reference-based guilt.

In principle there are, as illustrated in the introduction, clear differences between a belief-based and reference-based model of guilt. In particular, if λ_B differs widely from δ then we might obtain different predictions on how much effort Brian would exert. For example, if $\delta < \lambda_B$ then Brian will have a larger

[6] Brian's guilt is related to Ann's *payoff* expectations and so, generally speaking, Brian does not feel guilt from doing a different action than he believed Ann expected unless this lowers Ann's payoff.

[7] An alternative, or more general, model of belief-based guilt, proposed by Khalmetski, Ockenfels and Werner (2015), allows that people get pleasure from surprising others. In this case we may observe a negative correlation between effort and second-order beliefs.

incentive to exert effort according to the belief-based than reference-based model. To progress further, however, we need to tie down an interpretation of the reference level. In the following I will interpret the reference level as representing some social norm, or Brian's interpretation of a norm. It is though useful to distinguish between two common interpretations of a social norm (Andrighetto et al. 2015).

One approach is to interpret a norm as an *empirical expectation* of what others will do. In short, a norm becomes a norm when lots of people behave consistently with the norm (e.g. Schelling 1978, Sugden 1986). In this case the reference level of Brian should capture his expectation of what the 'average' person in his position would do. Specifically, we could think of δ as Brian's belief about the average effort chosen by workers. A second approach to interpreting social norms is to view them as a normative expectation of what people should do. With this interpretation, it is desirable to conform to the norm irrespective of whether others do or not (e.g. Bernheim 1994, Bicchieri 2006). The reference level δ would then represent Brian's *normative beliefs* about what is the right thing to do. This gives us two different versions of the reference-based model to compare.

Before we compare in detail the differences between the belief-based and reference-based model I want to briefly highlight the ways in which different incentives might interact. The possibilities here are endless and so I will just provide one illustrative example. Suppose that Brian has belief-based guilt but does not feel guilt over 'excessive' expectations of Ann (Khalmetski 2016). This requires a notion of excessive and so we need some form of reference level (which brings us to the reference-based model). One example is the following utility function

$$u_h(e, \lambda_B, \delta) = \begin{cases} -c(e) - \gamma_{bb} \max\left\{\pi\left(\lambda_B\right) - \pi\left(e\right), 0\right\} & \text{if } \lambda_B \leq \alpha\delta \\ -c(e) - \gamma_{sr} \max\left\{\pi\left(\delta\right) - \pi\left(e\right), 0\right\} & \text{otherwise} \end{cases} \tag{3}$$

where $\alpha \geq 1$ is a parameter. In interpretation, if Brian believes Ann's expectations are within factor α of his reference level then he experiences belief-based guilt, but if her expectations are excessive then he reverts to reference-based guilt. If $\alpha = 1$ then any expectation above Brian's reference level is considered excessive and so Brian always takes the 'most convenient' interpretation when judging guilt. If α is large than we have belief-based guilt.

Another thing to consider is heterogeneity of preferences. Some people may be motivated by belief-based guilt, others by empirical conformity and so on. Moreover, we may obtain complex interaction effects. For instance, the guilt from deviating from a normative belief may depend on the number of others deviating (Lopez-Perez 2008, 2012); hence, normative beliefs and empirical conformity become entangled. Or it may be that enough people with normative beliefs choose the same action that this then influences the second-order beliefs and empirical expectations of people not motivated directly by normative beliefs. Again, different motives to avoid guilt become entangled. This hints at the difficulties of comparing between models of guilt.

2.2 Comparing between models

The preceding discussion illustrates that we should not imagine there is a simple choice between the belief-based and reference-based models. Behavior will inevitable be influenced by a multitude of factors. Even so it is natural to try and unpick which factors seem most relevant. And in principle we now have three different things than can influence Brian's actions - his beliefs about what Ann expects, his empirical expectation as to what the average person does, and his belief about what he ought to do. The following hypothesis summarizes the analysis so far.

Hypothesis 1: With belief-based guilt aversion Brian's effort level is increasing in his second-order belief of Ann's expectations. With guilt based on empirical conformity Brian's effort level is increasing in his belief on average effort in the population. With guilt based on normative beliefs Brian's effort level is increasing in his normative belief.

If we elicit Brian's second-order beliefs, his belief on average effort in the population and normative beliefs then we can, in principle, apply Hypothesis 1 and pick apart different influences. Unfortunately, things are unlikely to be so simple in practice because of two distinct 'problems' - correlation of beliefs and the false consensus effect. Let us consider each in turn.

First, compare an empirically based reference level, δ, with Brian's second-order beliefs, λ_B. If Brian and Ann know nothing about each other then we should expect that $\delta = \lambda_B$. In particular, it seems reasonable that Brian would expect Ann to expect him to do what the average person in his position would do. But that means that second-order beliefs coincide with the empirically based reference level and we have no way to pick apart belief-based guilt and empirical conformity. With a normatively based reference amount there could be a clearer distinction between what Brian thinks he ought to do and what he believes Ann expects him to do. But even here it would not be a surprise to find a high correlation between normative beliefs and empirical expectations. For instance, if most people do what they think they ought to do then we should get a high correlation (Lopez-Perez 2008).

If Brian and Ann have specific knowledge about each other then things become more interesting because we might expect a divergence between δ and λ_B. Suppose, for instance, that Brian has a reputation for being a very hard worker. Then we might expect that $\lambda_B > \delta$ because Brian expects that Ann expects him to work harder than the average. Or if Brian has a reputation as a slacker then we might expect that $\lambda_B < \delta$. In such situations the belief-based model predicts that Brian focuses on the *expectation of Ann when interacting with Brian* while the reference-based model predicts that Brian focuses on an *average interaction between worker and employer*. This gives us something to go on, although, as we shall see in Section 4.2, the fact Brian and Ann knowing something about each other may bring in confounding factors.[8]

Another complicating factor is the false consensus effect. The false consensus effect captures the tendency of an individual to think that others are like them (Ross, Greene, and House 1977, Marks and Miller 1987, Engelmann and Strobel 2012). This may create a 'spurious' correlation between behavior and second-order beliefs (Ellingsen et al. 2010). The specifics of this will depend on the mechanism used to elicit beliefs (as discussed in Section 3). But to understand the basic point suppose, by way of argument, that the normative-based model is the 'true model'. Would we be able to pick this up? If Brian has a relatively high normative belief then he will exert high effort. The false consensus effect would then suggest that, because he exerts relatively high effort, he expects others to exert high effort. This, in turn, might result in him putting a high estimate on the effort Ann expects him to exert, and on average effort in the population.[9] We obtain, therefore, not

[8] If hybrid models are considered it also becomes difficult to disentangle the basis for guilt. Suppose, for example, that $\delta << \lambda_B$ and so there is a big gap between Brian's reference level and his belief on the expectations of Ann. This is precisely the kind of gap that should allow us to identify a difference between belief-based and reference-based guilt. If, however, Brian considers the λ_B belief as so outrageous that he can ignore it without feeling guilt, then we may prematurely rule out belief-based guilt.

[9] Whether this is 'false' is open to debate. As Dawes (1989) points out, it may be reasonable that person learns from their own choice. It only becomes 'false' if a person

only a ‘true’ positive correlation between effort and normative belief but also a spurious correlation between effort and second-order beliefs and between effort and empirical expectations. This means we may accept the belief-based model even though the normative-based model is the true one. Generally speaking, however, it is important to recognize that the false-consensus effect is not evidence against belief-based guilt aversion; it just means that care is needed in interpreting correlations between effort and beliefs.

If there is a high correlation between second-order beliefs, empirical expectations and normative beliefs (whether ‘real’ or caused by a false-consensus effect) Hypothesis 1 loses bite. In particular, a positive correlation between, say, effort and second-order beliefs is not compelling evidence in favor of belief-based guilt aversion. Crucially this means, while it is somewhat preempting the analysis to follow, that distinguishing between different models is going to be very difficult. To merely look for correlations between effort and beliefs is not enough. Moreover, any kind of treatment effect designed to shift beliefs is likely to shift all beliefs (second-order, empirical and normative) and so is also not enough. We will, therefore, need to look for relatively subtle ways to try and distinguish between models.

3 Beliefs and actions in the lab

Having set out a basic theoretical framework with which to work I shall now turn to a survey of experimental results. In doing so I shall pay particular attention to the setting considered by Charness and Dufwenberg (2006), namely a trust game with hidden action. This seems appropriate given that their benchmark game has been studied very widely and so we have a relatively large body of evidence with which to discern robust findings. Where appropriate I shall also bring in related evidence from studies of the dictator and trust game. Let me clarify, however, that I shall not delve deeply into the large literature on lie aversion (e.g. López-Pérez and Spiegelman 2013, Khalmetski 2016).

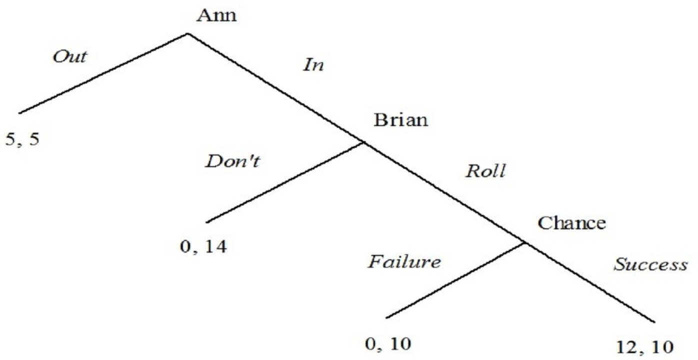
The benchmark (5,5) game studied by Charness and Dufwenberg (2006) is depicted in Figure 1. Ann, the sender, can choose either In or Out. If she chooses Out then the game ends with each getting a payoff of 5. If she chooses In then Brian, the receiver, can either Roll a dice or Don’t roll. If he does not roll then Ann gets a payoff of zero and Brian a payoff of 14. If the

overweights their own choice relative to other available evidence.

dice is rolled then Brian gets a payoff of 10 while the payoff of Ann depends on chance. There is a 1/6 chance of failure with Ann getting payoff of zero and a 5/6 chance of success with Ann getting a payoff of 12. Note that Ann is not informed whether a zero payoff is because Brian chose Don't or chance led to Failure. Charness and Dufwenberg (2006) also considered a (7,7) game where the payoff from Out is 7 for each player.

If we focus on monetary payoffs, in the (5,5) game, then a risk neutral Ann should choose In if and only if she believes there is a 50% or better chance that Brian will choose Roll. In the (7,7) game the relevant proportion is 70%. In general, Ann is only going to choose In if she believes there is a significant chance that Brian will choose Roll.[10] If, therefore, Ann chooses In and Brian chooses Don't then Brian can experience guilt because he 'lets down' Ann. Anticipation of this guilt may be sufficient for Brian to choose Roll. In relating this game to the theoretical example in Section 2 we can equate Roll with Brian choosing a high level of effort and Don't as choosing a low level of effort.

Figure 1: The (5,5) game of Charness and Dufwenberg (2006).



[10] If we add in social preferences then the belief necessary to incentivize Ann to choose In may well need to be more than 50% because the (In, Don't) outcome is 'bad' in that it is unkind, leads to inequality etc. Note also that Ann may suffer guilt from choosing Out if she believes that Brian is expecting her to choose In (Attanasi, Battigalli and Manzoni 2015).

Table 1: The average guess of receivers who choose Don't roll and Roll for the five treatments of Charness and Dufwenberg (2006) and directly comparable treatments in other studies.

Study	Treatment	Don't	Roll
Charness and Dufwenberg (2006)	(5,5) no message	39.6	54.2
	(5,5) B message	45.1	73.2
	(5,5) A message	50.0	69.6
	(7,7) no message	41.7	69.4
	(7,7) B message	36.9	66.9
Charness and Dufwenberg (2010)	(5,5) Bare promise	53.1	60.8
Ellingsen et al. (2010)	(5,5) no message	51.4	65.5
Amdur and Schmik (2013)	(5,5) no message	48.6	40.2

3.1 Guess-the-average approach

A belief-based model of guilt aversion would predict that Brian is more likely to Roll the higher is his second-order belief on the probability of Roll. But how to measure second-order beliefs?[11] Charness and Dufwenberg (2006) take the following approach, which I shall call the *guess-the-average approach*. Senders were asked to predict what proportion of receivers (in the experimental session) would choose Roll.[12] This is a proxy for first-order beliefs. And receivers were asked to guess the average guess made by senders. This is a proxy for second-order beliefs. Consistent with belief-based guilt aversion a strong correlation was observed between receiver's choice and their guess of the average. To illustrate Table 1 details the average guess of receivers who chose Don't and those who chose Roll in the treatments of Charness and Dufwenberg (2006) and those from directly comparable studies. (Note that we will look in detail at the role of messages in Section 4.)

To quote Charness and Dufwenberg (p. 1589, 2006) 'In all five treatments, B's who chose Roll made significantly higher guesses about A's guesses than did B's who chose Don't Roll. ... We conclude that the support for guilt

[11] A related question is whether elicitation of beliefs changes behavior. Guerra and Zizzo (2004) find evidence that it does not. But Ockenfels and Werner (2014) show that the way beliefs are elicited can have subtle effects.

[12] An experimental session would have many senders and receivers playing the (5,5) game in pairs.

aversion is considerable in all of our treatments.' It should be noted that Charness and Dufwenberg (2010), Ellingsen et al. (2010), and Andrighetto et al. (2015) do not find an effect that is statistically significant at the 5% level, and Amdur and Schmick (2013) actually observe a negative, again statistically insignificant, correlation.[13] If, however, we slightly extend our gaze to include studies with more substantive variations on the (5,5) game (e.g. Dufwenberg and Charness 2000, Chang et al. 2011, Bracht and Regner 2013, Bellemare et al. 2017) the evidence clearly points towards the following result.

Finding 1: There is a strong positive correlation between behavior and second order beliefs elicited using the guess-the-average approach.

A positive correlation between receiver's choice and guess is clearly consistent with the belief-based model. Is it also consistent with reference-based guilt? The guess-the-average approach provides a very natural estimate of beliefs regarding population averages. Finding 1, therefore, is also consistent with empirical conformity. In terms of normative beliefs, it seems reasonable to assume that a person can have one of two normative reference points, namely (Out, Don't) or (In, Roll). If Brian has normative expectation (Out, Don't) then he need feel no guilt by choosing Don't; Ann was 'dumb' to choose In. By contrast, if Brian has expectation (In, Roll) then he would feel guilt if he was to choose Don't. The normative beliefs model would, thus, predict that Brian is more likely to Roll if he thinks of (In, Roll) as the relevant reference point. There is no reason to suppose that the guess-the-average approach is a good measure of normative beliefs. A false consensus effect would, however, result in those who think of (In, Roll), or (Out, Don't), as the relevant reference point expecting others to also think that way. We may, therefore, expect a positive correlation between receiver's choice of Roll and guess as to what senders expect.

In summary, Finding 1 is consistent with belief-based guilt aversion and empirical conformity. If we allow for a false-consensus effect, then it is also consistent with normatively-based guilt. So, while Finding 1 is important evidence that guilt influences actions, it does not help us much distinguish between models of guilt.

[13] The results of Andrighetto et al. (2015) are not included in Table 1 because they do not provide data in this format.

3.2 The payoff-expectation approach

The guess-the-average approach is only one possible way of eliciting beliefs and various alternatives have been considered in the literature (see Schotter and Trevino 2014 for a general review). Most attention has focused on the disclosure approach that I shall discuss in the next section. Before doing that I briefly want to comment on the approach of Ismayilov and Potters (2016), which is closely related to that of Vanberg (2008) (see also Bacharach et al. 2007).

Ismayilov and Potters (2016) asked senders to say what they thought their payoff would be if they chose In. The options were: almost certainly 0, probably 0, not sure, probably 12, almost certainly 12. Receivers were then asked to guess which option the sender had chosen. Note that this approach differs from the guess-the-average approach in that it asks receivers to make a direct prediction on the payoff expectation of the *sender with whom they are matched* rather than predict population averages. For this reason I will call it the *payoff-expectation approach* to eliciting second-order beliefs.

If the receiver has not interacted with the sender in any way then we would expect the payoff-expectation approach to give similar results to the guess-the-average approach. In particular, the lack of contact means the receiver can only assume the sender with whom he is matched is 'average'. If, however, the sender and receiver do communicate then clearly the relationship becomes more personalized and so the payoff-expectation approach may yield different results to the guess-the-average approach. Indeed, the payoff-expectation approach would seem to more naturally capture the belief-based model's assumption that Brian cares about the expectations of Ann while the guess-the-average approach better captures the empirically conformity model's assumption that Brian cares about average behavior.

A comparison of the payoff-expectation and guess-the-average approaches would be of interest. Unfortunately, we do not have any evidence to go on. All we can say is that both Vanberg (2008) and Ismayilov and Potters (2016) find a positive correlation between the choice to Roll and second-order beliefs measured using the payoff-expectation approach. Ismayilov and Potters (2012), for instance, report that in the absence of communication those who Roll have average beliefs of 59% and those who Don't have average beliefs of 43%.[14] The corresponding beliefs in the case of communication are

[14] To put a number to average beliefs the 5 possible options are converted into percentages. For instance, almost certainly 0 is interpreted as an 87.5% probability of payoff 0

60% and 46%. This difference is highly significant and consistent with the numbers in Table 1. So, there is evidence that Finding 1 can be extended to include the payoff-expectation approach.

3.3 The disclosure approach

The *disclosure approach* to measuring, or perhaps more correctly inducing, second-order beliefs works as follows. Senders, as in the guess-the-average approach, are asked what proportion of receivers they expect will Roll. The receiver is then told the guess made by the sender, i.e. the receiver is told the first-order beliefs of the sender. In principle this means that receivers have 'perfect' second order beliefs because they know exactly what the sender expected. Note that there are some practical difficulties in implementing this approach because it is important that disclosure should not become a means of communication. Senders, therefore, should not be aware that their guess will be passed onto the receiver and receivers need to know that senders were not aware their guess would be passed on. This may lead to receivers questioning whether they have been 'told everything'.[15] Let us, however, put aside this issue here and focus on the results.

Ellingsen et al. (2010) and Kawagoe and Narita (2014) use the disclosure approach. Table 2 details the average guess observed by receivers who chose Don't and those who chose Roll. These results provide a stark contrast with those in Table 1. Crucially, there is no evidence that receiver's choice of Roll is correlated with the sender's guess (although Reuben, Sapienza and Zingales (2009) provide some contra-evidence in a trust game and Ockenfels and Werner (2014) for a dictator game).

Finding 2: The choice to Roll does not correlate with disclosed beliefs of senders.

The belief-based model would suggest a correlation between choice and

and probably 0 a 67.5% chance etc.

[15] There are different approaches to tackling this issue. Ellingsen et al. (2010) explicitly tell receivers that senders were not informed their guess would be passed on. Bellemare et al. (2017) do not explicitly tell receivers that senders were not informed. Reuben et al. (2009) use a multi-task experiment in which beliefs from one task are used in a subsequent task. This has the advantage that all subjects receive identical instructions. Khalmetski et al. (2015), see their experiment 3 and 4, allow senders to choose whether or not heir beliefs are passed to the receiver.

Table 2: The disclosed beliefs for receivers who choose Don't roll and Roll.

Study	Treatment	Don't	Roll
Ellingsen et al. (2010)	(5,5) No message	45.7	41.6
Kawagoe and Narita (2014)	(5,5) No message	51.7	44.5
	(5,5) B message	54.3	51.8

disclosed beliefs. Finding 2 is, therefore, fairly strong evidence against the belief-based model. Moreover, Ellingsen et al. (2010) obtain similar findings in two other games (a dictator game and trust game) and so the evidence against the belief-based model cannot be easily dismissed. At face-value this gives us two basic ways to interpret the 'discrepancy' between Findings 1 and 2. One is to claim it provides strong support for the reference-based model. In particular, it seems reasonable that the disclosed beliefs of the sender would not change the receiver's view on average behavior in the population or change their normative beliefs (because it is just the opinion of one other person). Seen in this light, Finding 1, which provides strong evidence of a correlation between choice and empirical expectations, together with Finding 2, no correlation with disclosed beliefs, points to the role of empirical conformity.

Another interpretation of the 'discrepancy' between Findings 1 and 2 is to say that Finding 1 is entirely driven by the false-consensus effect. If true then this would mean that guilt, whether belief-based or reference-based, is not a big influence on behavior. To appreciate the argument consider an extreme case in which heterogeneity of receivers' behavior is driven entirely by a misunderstanding of the instructions. The false consensus effect would then result in those who 'randomly' chose Roll expecting others to also choose Roll. This gives Finding 1. While they are not influenced by the disclosed belief of the sender. This gives us Finding 2. This is an extreme scenario but illustrates the challenge that Finding 2 provides.

So, is there any way to reconcile belief-based guilt aversion with Finding 2? For further insight consider the studies of Bellemare, Sebald and Strobel (2011) and Bellemare, Sebald and Suetens (2017). While neither study directly considers the (5,5) game of Charness and Dufwenberg they do study closely related dictator and trust games. Moreover, payoff parameters are systematically varied so as to estimate guilt sensitivity, or the willingness to pay to avoid guilt. The crucial thing for us, is that both studies directly compare the guess-the-average approach of eliciting second-order beliefs with the disclosure approach. Bellemare et al. (2017) also considers a menu approach that I shall look at in the next section.

Bellemare et al. (2011) estimate a structural model based on Battigalli and Dufwenberg (2007). In doing so they disentangle the false consensus effect from belief-based guilt aversion. Specifically, they use the guess-the-average approach to elicit beliefs before jointly estimating second-order beliefs and choices allowing for a correlation between guilt aversion and stated beliefs. This then allows them to say how much of the willingness to pay to avoid guilt is due to 'genuine' guilt aversion and how much is due to the correlation between guilt aversion and stated beliefs (i.e. the false consensus effect). Within subject comparison of treatments using the guess-the-average approach and disclosure approach then allows them to further pinpoint the size of the false-consensus effect. They find evidence of a large false-consensus effect (that increases estimated willingness to pay by factor 3) but also find, controlling for the consensus effect, significant evidence of a correlation between choice and second-order beliefs.

Even more telling is arguably the results of Bellemare et al. (2017). They find a strong correlation between choice and second-order beliefs using both the guess-the-average and menu approach. This is further evidence in support of Finding 1. They also found no correlation between choice and disclosed first-order beliefs. This is further evidence in support of Finding 2. Crucially, there design allows insight on why the disclosure approach gives different results and they find that the disclosure approach results in more 'cooperative' behavior (or to map into our context a higher proportion choosing Roll). In other words, a person appears more likely to Roll when the disclosure approach is used than when the guess-the-average approach is used.[16]

The results of Bellemare et al. (2017) suggest that the disclosure approach may not be an innocuous way of inducing beliefs. A likely explanation comes from the possibility that disclosing beliefs creates a connection between the sender and receiver. Note that experiments using the disclosure approach are carefully designed so that disclosure cannot be a form of cheap talk (or be perceived as a form of cheap talk). Even so, it still leads to the responder knowing something about the sender. As we shall see in Section

[16] This eliminates the correlation between choice and beliefs because it is those with low beliefs that are least likely to Roll (in the guess-the-average approach) and so more likely to be influenced.

4, a connection between sender and responder, however weak it may seem, can lead to an increase in the proportion who Roll. And it does not seem implausible that this effect would primarily impact on sender's with low expectations. This may offer an explanation for Finding 2. In the next sub-section we look at a related explanation.

3.4 Menu approach

With the *menu approach* Brian is able to condition his actions on the beliefs of Ann. More specifically, senders, as in the disclosure and guess-the-average approach, are asked what proportion of receivers they expect to Roll. The receiver is then asked to say whether or not he would Roll for all of the possible beliefs that Ann may disclose. Outcomes are then determined by Ann's disclosed belief and Brian's conditional action. The menu approach shares the same difficulties as the disclosure approach in the sense that Brian should not think Ann's beliefs are a form of communication (Khalmetski et al. 2015, see also Attanasi, Battigalli and Nagel 2016). Crucially, it also shares the property that choice is based on the disclosed beliefs of Ann, it is just that disclosure happens after Brian has written down his strategy.

As already noted, Bellemare et al. (2017) obtain similar results using the menu approach to the guess-the-average approach, which are different to those using the disclosure approach. Generally speaking, decisions elicited under the strategy method (which in our case would be the menu approach) are similar to those elicited under the action method (which in our case would be disclosure) but that does not mean there are not exceptions (Brandts and Charness 2000, 2011). The results of Bellemare et al. (2017) suggest that we potentially have one of those exceptions. This may point towards an experimenter demand effect in which asking subjects to condition on disclosed beliefs influences behavior, i.e. the menu approach is 'biasing' choice. This interpretation, however, seems hard to reconcile with the similarity between the guess-the-average and menu approach. The alternative interpretation is that the disclosure method has an effect on the interaction between Ann and Brian beyond mere revelation of beliefs (Schotter and Trevino 2014).

The study of Khalmetski et al. (2015) allows additional insight. They revisit the dictator game setting of Ellingsen et al. (2010) but using the menu approach rather than a disclosure approach (see their experiments 1 and 3). Consistent with Finding 2, and the results of Ellingsen et al. (2010), they observe no correlation between behavior and disclosed beliefs. The crucial new insight, however, is to recognize that this aggregate level correlation masks interesting individual level behavior. In particular, Khalmetski et al (2015) find that over 50% of subjects correlate their choice with beliefs (see also Attanasi et al. 2016). Most behave consistent with belief-based guilt-aversion in that their is a positive correlation but a significant proportion of subjects exhibit a negative correlation. It is this heterogeneity of behavior that causes the lack of correlation at the aggregate level.

In summary, we have seen two reasons why Finding 2 may not be compelling evidence against the belief-based model of guilt aversion. First, the disclosure method itself may change the relationship between Ann and Brian. In this regard it is worth highlighting that Khalmetski et al. (2015) interpret a negative correlation between choices and beliefs as a desire to make surprising gifts. This broadly fits the idea that reduced social distance between Brian and Ann may increase the Roll rate of Brian. A second reason to question the importance of Finding 2 is the evidence that interesting things are happening at the individual level even if this does not show through in aggregate level data. This is not to say the false-consensus effect does not exist; it clearly does. Indeed, Khalmetski et al. (2015) also elicit second order beliefs using the guess-the-average approach and reaffirm that there is evidence of a false-consensus effect. But, the weight of evidence suggests that the false-consensus effect is not responsible for *all* of the correlation between choice and second-order beliefs.

Finding 3: There is a correlation between choice and disclosed second-order beliefs using the menu approach. At the individual level this correlation may be positive (consistent with guilt) or negative (consistent with surprising gifts).

3.5 Normative beliefs

The guess-the-average approach to eliciting second-order beliefs fits naturally with empirically-based guilt aversion while the disclosure and menu approaches fit naturally with belief-based guilt aversion. As of yet, I have had little to say about normative beliefs. That is primarily because studies have not elicited normative beliefs.

Andrighetto et al. (2015) is an important exception. As well as eliciting second-order beliefs using the guess-the-average approach they elicit a range of normative beliefs. Specifically, senders were asked to say whether they feel entitled that Brian choose Roll. Receivers were asked whether they ought to choose Roll (own normative belief), to guess the proportion of senders who felt entitled (second-order normative expectations on senders), and to guess the proportion of receivers who feel they ought to Roll (second-order normative expectations on receivers). The headline result of Andrighetto et al. (2015) is that the choice to Roll (and Exit which we shall look at shortly) correlates significantly with the normative beliefs of the receiver (however these are measured).

Finding 4: There is evidence that choice to Roll correlates with normative beliefs.

Andrighetto et al. (2015) interpret their results as supporting the importance of normative beliefs over empirical conformity. Recall, however, that the false consensus effect makes it difficult to conclude too much from Finding 4, particularly given the evidence of a strong correlation between behavior and second-order beliefs. So, this is a topic that warrants further investigation.

4 Hidden actions and communication

The preceding section has illustrated that Hypothesis 1 is limited in application. The experimental evidence has given us a much better picture of how choices are influenced by beliefs and the likely size of the false-consensus effect. And we have fairly compelling evidence that guilt influences behavior in some form. But, whether that guilt is belief or reference-based remains unclear. This suggests we may have to look beyond correlation between choices and beliefs if we want to disentangle different influences. Two possibilities about which some evidence exists are the observability of actions and communication.

4.1 Hidden actions

In introducing the example in Section 2 it was left ambiguous whether Ann would observe the effort of Brian (or be able to discern the reason why her payoff is as it is).[17] Suppose that Ann will not know how much effort Brian

[17] Recall that $\pi(e)$ is expected payoff and so there is scope for random shocks etc.

exerted because, say, output is subject to random shocks or quality of service is unobservable. Both Charness and Dufwenberg (2006) and Battigalli and Dufwenberg (2007) allow that a person can experience *simple* belief-based guilt when actions are not observable. There are though strong arguments, that because of *guilt from blame*, belief-based guilt will be higher when effort is observed (Charness and Dufwenberg 2011).[18]

In the reference-based model we could also distinguish simple guilt and guilt from blame. The nature of both empirical conformity and normatively based guilt suggest, however, that Brian is not interested, per-se, in whether he lets Ann down. More relevant would seem to be whether effort is *publicly* observable. A nice combination of treatments in the study by Tadelis (2011) illustrates the issue. In one treatment there is matched pairs and public exposure, meaning that everyone in the room is told the action of Brian and Ann knows that she was matched with Brian. Here belief-based guilt from blame should be prominent because Brain knows that Ann will know what he did. In another treatment there is anonymous public exposure, meaning that everyone knows what Brian did but now Ann cannot know for sure she was matched with this particular Brian. Here, belief-based guilt from blame seems less relevant but reference-based guilt would seem unaffected.

Variations in exposure provide a way to distinguish variants of the belief-based model, such as, simple guilt and guilt from blame. And, as we have just seen, may also provide a way to distinguish between belief-based and reference-based guilt. The following hypothesis picks up some elements of this.

Hypothesis 2: With simple belief-based or reference-based guilt the effort of Brian is not influenced by the observability of his actions. With belief-based guilt from blame the effort of Brian is influenced by the observability of his actions by Ann. With reference-based guilt from blame the effort of Brian is influenced by whether his actions are publicly observable.

Now to the evidence. In the original Charness and Dufwenberg (2006) design, senders are not told why they get a payoff of 0. Brian, therefore, acts knowing that Ann will not be able to tell if he chose Roll, and she

[18] There is also a distinction between outcome disappointment where Brian feels guilt if he expects Ann to get a lower payoff than she expected and person disappointment where Brian feels guilt if he expects Ann to form a negative opinion of him (Bacharach et al. 2007).

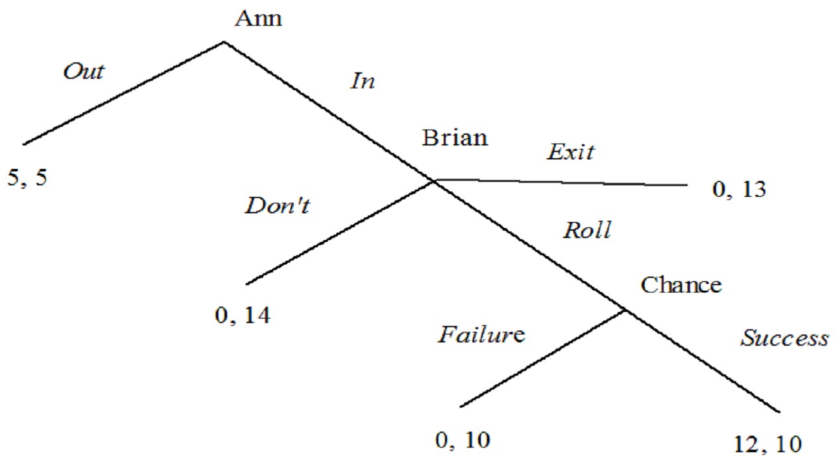
was unlucky, or he chose Don't, and caused her low payoff. This lack of exposure to the cause of a low payoff puts the focus on simple guilt (Battigalli and Dufwenberg 2007). If Ann is able to observe, or is exposed to, the actions of Brian then we have a setting where guilt from blame is possible. A comparison of treatments with and without exposure allows us, therefore, to gain insight on Hypothesis 2.

Three studies have looked specifically at exposure. Both Bracht and Regner (2013) and Tadelis (2011) find that a significantly higher proportion of receivers chose Roll when there is exposure. Andrighetto et al. (2015) do not find a statistically significant effect when directly comparing treatments with and without exposure. Their study does, however, nicely demonstrate the importance of exposure. Specifically, they consider a game, see Figure 2, in which Brian can choose Exit. If Brian chooses Exit then Ann is told that Brian chose Roll. So, Brian can effectively pay to escape any guilt from blame. Around 20% of subjects chose Exit. This shows that exposure matters. Note, however, that 17 to 32% of subjects (depending on the treatment) still chose to Roll, even though they could Exit. These results point towards a role for both simple guilt and guilt from blame. For instance, those choosing Exit avoid guilt from blame while those choosing Roll, rather than Exit, avoid simple guilt.

Recall that Tadelis (2011) compares various treatments combining matched exposure (where the sender is informed) to those with public exposure (where actions are announced to everyone in the room). He finds that exposure whether public or merely to the sender has an almost identical effect on behavior. This would seem something of a challenge to the reference-based model. To defend the model we would need to argue that Brian equates Ann observing his action with his action being publicly observable. Or that Brian does not care directly about Ann's payoff expectations but does care about whether Ann observes him disappointing those payoff expectations. This is something that could be explored in future work.

Finding 5: The decision to Roll is influenced by whether actions will be observable to the sender or publicly observable. The type of exposure appears to make no difference.

Figure 2: The exit game of Andrighetto et al. (2015).



4.2 Communication from Brian to Ann

Much of the literature on guilt aversion has focused on communication. So, returning to the example in Section 2, suppose that Brian makes a promise to Ann that he will exert effort level p. This promise may change Brian's beliefs about the effort level Ann is expecting. We do need to consider the credibility of the promise p, because if Brian has no intention of keeping his promise then why should he expect Ann to believe his promise. Even so, there is a clear sense in which beliefs, λ_B, should move towards p.

What of the reference-based model? As the reviewer of an earlier version of the paper pointed out, we have to condition empirical expectations and normative beliefs on the fact that a promise has been made. For instance, the reference level for empirical conformity would be based on average behavior of those who have promised to exert effort level p. In the normative-based model it would be based on what is the right thing to do given a promise of p. This suggests that a promise can also change the reference level (Lopez-Perez 2012). And again we would expect it to change in the direction of the promise.

The preceding discussion suggests that promises will likely matter. That is not a particularly bold hypothesis. Crucially, however, it may give us more to go in terms of Hypothesis 1. In particular, communication may start to drive a wedge between Brian's second-order beliefs and empirical expectations (or normative beliefs) because Ann and Brian start to 'know something about each other'. For that to work we ideally would like that promises do not have an effect beyond beliefs. So, an important hypothesis, following Charness and Dufwenberg (2006), is that the effect of promises is solely captured by changes in beliefs. In other words we can still apply equations (1) or (2) we just have to recognize that a promise may change beliefs, λ_B, or the reference level δ.

Hypothesis 3: A promise from Brian to Ann may change his behavior but *only* through the effect the promise has on second-order beliefs or the reference level.

Hypothesis 3 encapsulates a 'pure' form of belief-based guilt in that it says *only* beliefs matter. Similarly it encapsulates a pure form of reference-based guilt. So, what are the alternatives? It may be that promises have an influence beyond any effect on beliefs or the reference-level. For instance, Kawagoe and Narita (2014) argue that a person only feels guilt if they make a promise that is believed and then fail to fulfill the promise. With this approach letting someone down, of itself, does not induce guilt; which is a big step away from the belief-based and reference-based models outlined in Section 2. Instead, Brian only feels guilt if he promises to exert effort and then lets Ann down. The intuition here is that Brian feels guilt if he directly influences Ann's beliefs and then let's down those beliefs. The model proposed by Kawagoe and Narita (2014) is an extreme possibility that would struggle to explain the results in Section 3. But it nicely illustrates the more general point that Brian's guilt can be amplified by any change in Ann's payoff expectations that was induced by a promise he made (see also Balafoutas and Sutter 2016).

A related issue is that of lie-aversion. Lie aversion, at its most basic level, says that Brian will dislike lying (or failing to keep a promise).[19] This also means that promises may have an influence beyond their effect on beliefs. To illustrate, suppose, returning to our theoretical example, that Brian believes that Ann was expecting an effort level of 20. If he exerts effort 10

[19] In the simple setting considered in Section 2 the incentive to lie is not explicitly modeled. But, generally speaking, Brian has an incentive to lie if this will induce Ann to, say, pay a higher wage.

without making any promises then his payoff is $-c(10) - \gamma_{bb}\left(\pi\left(20\right) - \pi(10)\right)$ where the second term represents his guilt from disappointing Ann's *payoff expectation.* Now, suppose that Brian promises to exert effort level 20. Clearly this promise should not be expected to change Brian's belief about the effort Ann was expecting. It seems, though, perfectly reasonable that Brian also feels 'guilt from lying'. His payoff would now be something like $-c(10) - \gamma_{bb}\left(\pi\left(20\right) - \pi(10)\right) - G$ where the G term represents his guilt from *not fulfilling his promise.*[20] Guilt from lying, therefore, adds an extra influence on behavior.

In the benchmark treatment of Charness and Dufwenberg (2006) subjects could not communicate. This was compared with treatments where Brian is allowed to send a message to Ann and a treatment where Ann is allowed to send a message to Brian. Note that a free form message was allowed in which anything could be written. Table 3 summarizes the proportion of receivers choosing Roll by treatment. We can see that if Brian is able to send a message there is a significant increase in the proportion who choose Roll. A similar result is observed by Kawagoe and Narita (2014).[21] In interpreting these numbers it is important to take account of the message that was sent. Charness and Dufwenberg (2006) find that receivers who promised to Roll were significantly more likely to Roll than those who did not promise.

Overall, the results of Charness and Dufwenberg (2006) lend support to Hypothesis 3 in that promises matter (see Table 3) and influence beliefs (see Table 1). But that is not the end of the story. To further the discussion I will consider in some detail two papers that look at careful manipulations of communication.

Consider first the study of Vanberg (2008). Before we delve into the results let me highlight that the setting considered by Vanberg (2008) is somewhat different to that considered by Charness and Dufwenberg (2006). Vanberg (2008) essentially does away with the In or Out choice of Ann. So, we end up with a dictator game in which Brian chooses whether to Roll or Don't Roll. Another twist is that the two players communicate between each other before they know who will be in the role of Ann and who will

[20] There are many different models of lie aversion including belief-based lie aversion (Lopez-Perez and Spiegelman 2013). It is beyond the scope of the current paper to delve deeply into that issue. But it is worth highlighting that there is now abundant evidence people dislike lying per-se (Erat and Gneezy 2012).

[21] The effect observed by Kawagoe and Narita (2014) is not statistically significant (Fisher's exact test, $p = 0.415$) but the number of observations is small.

Table 3: The proportion (%) of receivers who choose Roll depending on the type of communication varying between no communication (no), communication from Brian to Ann (B), from Ann to Brian (A) and between Ann and Brian (AB).

Study	Treatment	No	B	A	AB
Charness and Dufwenberg (2006)	(5,5)	44	67	39	
	(7,7)	25	49	-	
Kawagoe and Narita (2014)	(5,5)	30	42	-	
Vanberg (2008)	Dictator	51	-	-	74
	No switch	-	-	-	69
	Switch	-	-	-	54
Ismayilov and Potters (2016)	Can promise	28	51		
	Cannot promise	45	52		
Bracht and Regner (2013)	(5,5)	33	56		
	Exposure	42	67		

be Brian. It is this uncertainty which creates the main incentive for making promises. In particular, a subject may promise he will Roll in the hope this would encourage the other subject to Roll if they end up being the dictator. Despite these differences Vanberg (2008) finds similar results to Charness and Dufwenberg (2008) in control treatments with and without communication (see the dictator row in Table 3).

The main novelty of Vanberg (2008) is to randomly switch the matching of subjects after communication has taken place. The dictator is aware if a switch has been made but the other subject is not. Moreover, the dictator is made aware if a promise was made to his new match. Vanberg (2008) finds that dictators were significantly more likely to Roll if they remained matched to the subject with whom they had communicated than if they were switched. Crucially, however, second-order beliefs (measured using the payoff-expectation approach) were the same whether a switch took place or not. This result is inconsistent with Hypothesis 3 and the Charness and Dufwenberg (2006) claim that behavior can be captured *entirely* through second-order beliefs. If this was the case then the Roll rate would be the same in the switch and no switch condition (given that second-order beliefs remain unchanged).

Consider next the study of Ismayilov and Potters (2016). The starting point here is the (5,5) game of Charness and Dufwenberg with Brian being able to send a message to Ann. The main twist is that there is only a 50% chance of a message being delivered. Brian is told whether or not the message was delivered before deciding to Roll. Ismayilov and Potters (2016) find that receivers who made a promise to Roll were significantly more likely to act on that promise if the message was delivered. More surprising, is that receivers who *did not make a promise* were also significantly more likely to Roll if the message was delivered. Indeed, receivers who sent a blank message were more likely to Roll when the message was delivered! Moreover, it is shown that this effect cannot be accounted for solely through changes in second-order beliefs.

It seems, therefore, that the mere act of communication seems to make a difference. A further treatment explores this by allowing Brian to send a message to Ann but not make any promises about how he will play the game. Ismayilov and Potters (2016) find that receivers are again more likely to Roll if the message was delivered, although the effect in this instance is statistically insignificant. Note that the fact communication, of itself, can have an effect is also consistent with the results of Vanberg (2008) where dictators are more likely to Roll if they communicated with the current match. We can summarize as follows.

Finding 6: The proportion of receivers who Roll is influenced by communication from Brian to Ann. But this influence cannot be accounted for solely through a change in second-order beliefs.

If we compare Finding 6 with Hypothesis 3 then the evidence on communication is mixed. So, what can we take away from this? There can be no doubt that the literature studying guilt and communication has taught us a lot about communication. I am less convinced that it has taught us much about guilt-aversion. The original hope of Charness and Dufwenberg (2006), that the effects of communication can be entirely captured with a belief-based model of guilt aversion, has proved overly optimistic. Instead, we have learned that communication appears to bring with it a host of confounding factors including lie-aversion and reduced social-distance. It particularly seems to be the case that communication interacts with guilt aversion in important ways (see also Kawagoe and Narita 2014 and Balafoutas and Sutter 2016). For instance, it may that Brian feels more guilt if he lets down expectations that were directly influenced by him.

Interestingly, even Finding 6 and the rejection of Hypothesis 3 leaves ample scope for belief-based guilt aversion to explain all the effects of communication. To appreciate this point consider Khalmetski (2016). Here guilt from lying is modeled using belief-based guilt aversion (see also Battigalli, Charness and Dufwenberg 2013). A clever experiment design is then used to induce an exogenous shift in second-order beliefs. Crudely put, two treatments were compared in which the amount a lie would let down the payoff expectations of a receiver were high or low. Belief-based guilt aversion then predicts second-order beliefs on the rate of honesty are larger in the high treatment. This, in turn, should lead to more truth-telling in the high treatment. Khalmetski (2016) finds strong support for the belief-based model. So, belief-based guilt aversion can explain why someone would keep a promise. We then need to consider the interaction between guilt from failing to repay trust and guilt from failing to keep a promise.

The extent to which communication can help us disentangle the evidence for belief-based guilt aversion appears, therefore, moot. Let me emphasize, however, that this should not be interpreted as evidence against guilt aversion. It merely reflects the fact that communication brings with it other factors that seem hard to control. Several studies have shown that second-order beliefs still matter when there is communication suggesting that the models of guilt discussed in Section 2 are still relevant (Charness and Dufwenberg 2006, Andrighetto et al. 2015, Ismayilov and Potters 2016).

4.3 Communication from Ann to Brian

Consider finally the possibility that Ann can communicate with Brian. This will presumably take the form of Ann suggesting how much effort Brian should exert. Such a suggestion may change Brian's beliefs about the effort level Ann is expecting. Again, the consequences of this in a belief-based model of guilt can already be captured in equation (1). Clearly, we need to consider the credibility of the suggestion. It seems reasonable, however, to assume that the suggestion can have a significant effect on Brian's second-order beliefs. If, for instance, Ann's suggestion is below the level that Brian was expecting then he will surely revise down his second-order beliefs.

In a reference-based model of guilt, by contrast, there is little reason why Ann's suggestion should change Brian's reference level. It is plausible that an 'unexpected' suggestion may make Brian revise his estimate of average effort in the population or question his normative beliefs. The effect, however, is likely to be small, particularly if we take account of the false consensus effect. This leads to our next hypothesis.

Hypothesis 4: With belief-based guilt the action of Brian will weakly move in the direction of Ann's suggestion. With reference-based guilt the action of Brian should not depend on the suggestion of Ann.

Crucially, lie aversion is not going to matter here. Communication from Ann to Brian is, therefore, potentially a good way of distinguishing between the belief-based and reference-based model. I will, however, mention one possible confound. Namely, it may be that Ann's suggestion changes Brian's intrinsic desire to exert effort. This could manifest itself in different ways. For instance, a relatively low suggestion may motivate Brian to surprise Ann with high effort (Khalmetski et al. 2015). Or, a high suggestion, even if it was consistent with Brian's beliefs, may appear presumptuous and crowd-out intrinsic motivation (Bacharach et al 2007). A suggestion may, therefore, have an effect in the reference-based model despite not changing the reference level because it changes other things like sensitivity to guilt. Or, more generally, that it brings in other factors like reduced social distance.

Summarizing the experimental results on communication from Ann to Brian is easy enough for the simple reason that Charness and Dufwenberg (2006) is the only study of which I am aware that considers this possibility. As you can see in Table 3, such communication appears to make no difference to Brian's choice. Applying Hypothesis 4 this is evidence in support of the reference-based model. Clearly, however, it would be nice to have more evidence on this issue.

5 Conclusion

My objective in this paper was to review the evidence on belief-based guilt aversion, with a particular focus on behavior in trust and dictator games. The belief-based model says that a person experiences guilt if they let down the payoff expectations of another (Battigalli and Dufwenberg 2007). The predictions of this model were compared to those of a reference-based model in which a person experiences guilt if they deviate from a norm, either empirically or normatively based. The theoretical analysis highlighted the difficulties of distinguishing between the different models. This carried through to the review of the experimental evidence where we find broad support for both models.

While the basic conclusion is that the existing experimental evidence supports both the belief-based and reference-based models of guilt aversion it is important to recognize that not all the evidence is positive. In particular, evidence that exposure, or observability of actions, makes a difference to behavior is easier to reconcile with the belief-based model than reference-based model (e.g. Tadelis 2011, Bracht and Regner 2014). On the other hand, evidence that behavior is more closely correlated with beliefs elicited using a guess-the-average approach than disclosure approach is easier to reconcile with the reference-based model than belief-based model (e.g. Ellingsen et al. 2010). The picture, therefore, is somewhat mixed.

One way to interpret these findings is to acknowledge the difficulty of distinguishing between different models of guilt. Just about all the evidence reviewed in this paper suggests that the belief-based model of guilt aversion can help make valid predictions on behavior, even if it, not unsurprisingly, does not capture everything. Moreover, there is more which unites than divides the belief-based and reference-based models. For instance, a key point made by Charness and Dufwenberg (2006, 2010) is that feelings of guilt are very much context dependent. This point easily translates to the reference-based model. Support for the reference-based model could, therefore, be seen as a critique or endorsement of the belief-based approach depending on one's perspective.[22]

Another thing to take from this review is the potential merits of a hybrid approach that incorporates aspects from both the belief-based and reference-based models. For instance, the belief-based model does not question whether beliefs are 'reasonable'. Yet, a person may feel no guilt from disappointing unreasonable payoff expectations (Khalmetski 2016). Or they may want to surprise someone with low expectations (Khalmetski et al. 2015). The difficulty with this interpretation is that we need a working notion of reasonable or low and that inevitable means we need to bring in some kind of reference level. Also, the belief-based model comes into it's own when the people involved know something about each other (and so have more informed beliefs), but we have seen that this reduced social distance brings with it complicating

[22] This conclusion differs from that obtained for other belief-based models, such as reciprocity, where predictions are clearly sensitive to the model used (e.g. Falk, Fehr and Fischbacher 2008).

effects such as increased cooperation. Note that this is not merely to say that belief-based guilt aversion is one of many factors that influence choice. But more a call for the need to model the interaction between different factors.

So, what direction could future work take? Here are some suggestions of things that could be built into or considered in experimental work.

- We need more evidence on normative beliefs and so there is merit in asking subjects for their normative beliefs alongside eliciting their second-order beliefs. Andrighetto et al. (2015) show the way forward in this regard. It would also be of interest to see how normative beliefs are influenced by, say, exposure or communication. For instance, does the 'right thing to do' change when Brian knows that Ann will observe his actions.

- We can elicit second-order beliefs using multiple different approaches for the *same* subject. There has possibly been a reluctance to use multiple approaches to elicit the 'same thing'. But from the perspective of subjects things like guess-the-average, payoff-expectation and disclosure approaches may feel different. Moreover, one can argue that they measure subtly difference things. For instance, the guess-the-average approach is about population averages while the payoff-expectation approach captures an element of uncertainty. So, there is no reason to not obtain data using multiple approaches. This, coupled with an elicitation of normative beliefs, is the simplest and most direct way of testing Hypothesis 1. Bellemare et al. (2011, 2017) show the way forward in this regard.

- Communication is a potential way to distinguish between models of guilt aversion. But, it also brings along many confounding factors. To control for that it would be beneficial to elicit second-order beliefs (and potentially normative beliefs) before and after communication. For instance, we could elicit Brian's second-order beliefs using the guess-the-average approach, allow him to send a promise to Ann, and then remeasure beliefs. Or we could use the strategy method to see how, say, Brian's second-order beliefs are influenced by a promise being received or not by Ann. This would move us away from trying to extrapolate treatment effects from point estimates of beliefs. We also see the change in beliefs.

- The literature has almost exclusively focused on communication from Brian to Ann. But communication from Ann to Brian seems equally interesting and so this is an issue that warrants future work. Also, a reviewer suggested that one could consider communication to Brian from an independent third-party. This may provide a very nice of way influencing Brian's beliefs without the confounding problem of communication between Ann and Brian. Again, we can measure beliefs in multiple ways before and after communication.

- Exposure is an issue that also seems understudied. Tadelis (2011) shows the way in comparing matched and public exposure with anonymity. But again it would be interesting to pick apart the consequences of exposure for beliefs, empirical expectations and normative beliefs to then better understand why exposure matters.

- There is merit in considering the diffuseness of beliefs. The standard model of belief-based guilt aversion assumes that higher-order moments of beliefs do not influence guilt (Batigalli and Dufwenberg 2007). Intuitively, however, the more uncertainty Brian has on Ann's beliefs, then the less guilt he may feel (Bacharach, Guerra and Zizzo 2007).[23] This can have several important implications. For instance, one effect of communication may be to reduce uncertainty. Also, the disclosure approach may differ from the menu approach because it reduces Brian's uncertainty. To capture uncertainty we need to elicit not only second-order beliefs but uncertainty around beliefs.

- The study of Khalmetski (2016) provides a beautiful illustration of how an exogenous shift in second-order beliefs can be induced through a careful manipulation of treatments. The focus of that study was on guilt from lying and so it cannot directly speak to the distinguish between belief-based and reference-based guilt in trust games. It does, though, show the merit of trying to induce an exogenous shift in beliefs.

A slightly bigger challenge comes with the application of guilt aversion. Theoretical elegance suggests an equilibrium approach in which beliefs are correct (Charness and Dufwenberg 2007, e.g. Beck et al. 2013). Reality, however, suggests that beliefs will be noisy and that people may not fully

[23] This would follow from convexity of π but also may reflect less guilt if there is wiggle room to justify actions (Dana, Weber, Kuang 2007).

anticipate guilt when making decisions. In particular, belief-based guilt aversion does not require that beliefs be correct. It merely requires that people act on their beliefs. This muddies the waters. Suppose, for instance, as often seems to be the case, that first-order beliefs are biased. This might lead to unreasonable expectations and almost certainly leads to biased second-order beliefs.

The consequences of this need to be carefully unpicked. Moreover, if guilt is an emotion someone experiences *after* an event then it may be unclear how much *anticipation* of guilt influences behavior (Miettinen and Suetens 2008, Bracht and Regner 2013). This points to a need for more understanding of how people form beliefs, how sensitive they are to guilt, and how anticipation of guilt feeds into decisions (Bellemare et al. 2011, Bracht and Regner 2013, Khalmetski et al. 2015 and Attanasi et al. 2016 are all important steps in this direction). For instance, a subject may walk out of the experimental lab feeling no guilt, talk to another subject, update his beliefs and then feel guilt for the decisions he made in the experiment. If people are poor at predicting (particularly in the lab) whether a choice will subsequently make them feel guilty, choices may not capture the full extent of guilt.

References

Amdur, D., & Schmick, E. (2013). Does the direct-response method induce guilt aversion in a trust game?. *Economics Bulletin*, 33(1), 687-693.

Andrighetto, G., Grieco, D., & Tummolini, L. (2015). Perceived legitimacy of normative expectations motivates compliance with social norms when nobody is watching. *Frontiers in psychology*, 6.

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

Attanasi, G., Battigalli, P., & Nagel, R. (2016). Disclosure of Belief-Dependent Preferences in a Trust Game. Working paper.

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

Balafoutas, L., & Sutter, M. (2016). On the nature of guilt aversion: Insights from a new methodology in the dictator game. *Journal of Behavioral and Experimental Finance.*

Battigalli, P., & Dufwenberg, M. (2007). Guilt in games. *The American Economic Review*, 97(2), 170-176.

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

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

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

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

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

Bernheim, B. D. (1994). A theory of conformity. *Journal of political Economy*, 102(5), 841-877.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms.* Cambridge University Press.

Bracht, J., & Regner, T. (2013). Moral emotions and partnership. *Journal of Economic Psychology*, 39, 313-326.

Brandts, J., & Charness, G. (2000). Hot vs. cold: Sequential responses and preference stability in experimental games. *Experimental Economics*, 2(3), 227-238.

Brandts, J., & Charness, G. (2011). The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics*, 14(3), 375-398.

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

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

Charness, G., & Dufwenberg, M. (2010). Bare promises: An experiment. *Economics Letters*, 107(2), 281-283.

Charness, G., & Dufwenberg, M. (2011). Participation. *The American Economic Review*, 101(4), 1211-1237.

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

Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology*, 25(1), 1-17.

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

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

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

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

Engelmann, D., & Strobel, M. (2012). Deconstruction and reconstruction of an anomaly. *Games and Economic Behavior*, 76(2), 678-689.

Erat, S., & Gneezy, U. (2012). White lies. *Management Science*, 58(4), 723-733.

Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing theories of fairnessIntentions matter. Games and Economic Behavior, 62(1), 287-303.

Geanakoplos, J., Pearce, D., & Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, 1(1), 60-79.

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

Huang, P. H., & Wu, H. M. (1994). More order without more law: A theory of social norms and organizational cultures. Journal of Law Economics and Organization, 10, 390.

Ismayilov, H., & Potters, J. (2016). Why do promises affect trustworthiness, or do they?. *Experimental Economics*, 19(2), 382-393.

Ismayilov, H & Potters, J. (2012). Promises as Commitments. CentER Discussion Paper 2012-064.

Kawagoe, T., & Narita, Y. (2014). Guilt aversion revisited: An experimental test of a new model. *Journal of Economic Behavior & Organization*, 102, 1-9.

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

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

Lpez-Prez, R. (2012). The power of words: A model of honesty and fairness. *Journal of Economic Psychology*, 33(3), 642-658.

López-Pérez, R. (2010). Guilt and shame: An axiomatic analysis. *Theory and Decision*, 69(4), 569-586.

Lpez-Prez, R. (2008). Aversion to norm-breaking: A model. *Games and Economic Behavior*, 64(1), 237-267.

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

Marks, G., & Miller, N. (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin*, 102(1), 72.

Miettinen, T., & Suetens, S. (2008). Communication and Guilt in a Prisoner's Dilemma. *Journal of Conflict Resolution*, 52(6), 945-960.

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

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

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

Schelling, T. C. (2006). *Micromotives and macrobehavior*. WW Norton & Company.

Schotter, A., & Trevino, I. (2014). Belief elicitation in the laboratory. *Annual Review Economics*, 6(1), 103-128.

Sugden, R. (1986). *The economics of rights, co-operation and welfare*. Oxford: Basil Blackwell.

Tadelis, S. (2011). The power of shame and the rationality of trust. Haas School of Business working paper.

Vrij, A. (2000). *Detecting Lies and Deceit: The Psychology of Lying and Implications for Professional Practice*. Wiley.