Cognitive Psychology xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Seeing is believing: Trustworthiness as a dynamic belief

Luke J. Chang [a], Bradley B. Doll [b], Mascha van 't Wout [b], Michael J. Frank [b], Alan G. Sanfey [a,*]

[a] Department of Psychology, University of Arizona, 1503 E. University Blvd, Tucson, AZ 85721, United States
[b] Departments of Cognitive & Linguistic Sciences and Psychology, Brown University, 190 Thayer St, Providence, RI 02912-1978, United States

ARTICLE INFO

Article history:
Accepted 22 March 2010
Available online xxxx

Keywords:
Decision-making
Trust
Social learning
Trustworthiness
Reinforcement learning
Face perception

ABSTRACT

Recent efforts to understand the mechanisms underlying human cooperation have focused on the notion of trust, with research illustrating that both initial impressions and previous interactions impact the amount of trust people place in a partner. Less is known, however, about how these two types of information interact in iterated exchanges. The present study examined how implicit initial trustworthiness information interacts with experienced trustworthiness in a repeated Trust Game. Consistent with our hypotheses, these two factors reliably influence behavior both independently and synergistically, in terms of how much money players were willing to entrust to their partner and also in their post-game subjective ratings of trustworthiness. To further understand this interaction, we used Reinforcement Learning models to test several distinct processing hypotheses. These results suggest that trustworthiness is a belief about probability of reciprocation based initially on implicit judgments, and then dynamically updated based on experiences. This study provides a novel quantitative framework to conceptualize the notion of trustworthiness.

1. Introduction

The success of human civilizations can be largely attributed to our remarkable ability to cooperate with other agents. Cooperative relationships, in which individuals often endure considerable risk, are built on the foundation of trust – a nebulous construct that is nevertheless intimately tied to both

* Corresponding author. Fax: +1 520 621 9306.
E-mail address: asanfey@u.arizona.edu (A.G. Sanfey).

0010-0285/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.cogpsych.2010.03.001

interpersonal (Rempel, Holmes, & Zanna, 1985) and economic prosperity (Zak & Knack, 2001). However, as anyone who has ever purchased a used car can attest, not everyone turns out to actually be trustworthy. Thus, the accurate inference of an individual's level of trustworthiness is crucial for the development of a successful relationship. How then does one actually assess trustworthiness? Business people often attest to the importance of looking a future partner in the eye and physically shaking their hand before signing a contract. When physical meetings are not an option, people frequently rely on reputation, which in the world of online commerce has taken the form of buyer testimonials about a seller's prior transactions on sites such as EBay. This is consistent with the notion that the best predictor of an individual's level of trustworthiness is their behavior in previous interactions with us (Axelrod & Hamilton, 1981; King-Casas et al., 2005). We are more likely to invest trust in someone previously shown to be trustworthy than someone who has previously betrayed us. Therefore, one useful model for inferring the trustworthiness of another is to make an initial assessment based on available information, and then update this judgment based on subsequent interactions.

Because trust is an amorphous construct, it is often difficult to measure and operationalize. From a psychological perspective, trust can be considered the degree to which an individual believes that a relationship partner will assist in attaining a specific interdependent goal (Simpson, 2007). While this definition can apply to any number of social interactions, consider an example of confiding in a colleague. Alex has been privately deliberating a decision and seeks feedback from Trevor. Alex is interested in Trevor's perspective, but does not want David to know. Trust, in this example is Alex's belief that Trevor will not divulge this sensitive information to David. Of course, trust is a broad concept that extends to many different aspects of social interaction and may depend on the assessment of a variety of factors, including, honesty, competence, competiveness, and greed. However, in order to study trust experimentally, it is necessary to have a good operationalization of it, even though this may be limiting.

The Trust Game is a task that has been developed by Behavioral Economists to serve as a proxy for everyday situations involving trust like in our example. The Trust Game explicitly measures our Psychological operationalization of trust, and assesses the degree to which an individual is willing to incur a financial risk with a partner (Berg, Dickhaut, & McCabe, 1995). This simple game involves two players, A and B. Player A is endowed with an initial amount of money, say $10, and can choose to invest any amount of this endowment with B. The amount that Player A invests is multiplied by the experimenter by some factor, usually 3 or 4, and then Player B decides how much of this enlarged endowment, if any, they would like to return to Player A. The partner can choose to repay the investor's trust by returning more money than was initially invested, or abuse their trust by keeping all (or most) of the money. In this game, trust is operationally defined as the amount of money that a player invests in their partner, and trustworthiness is defined as the likelihood that the partner will reciprocate trust. Evidence from empirical work has shown that most investors are willing to transfer about half of their endowment. In turn, when the investment is multiplied by a factor of 3, partners are usually willing to reciprocate trust so that both partners end up with approximately equal payoffs (Berg et al., 1995). This simple game provides a useful behavioral operationalization of trust, and also demonstrates that in general players exhibit both trust and trustworthiness, contrary to the standard predictions of economic Game Theory (Camerer, 2003). Additionally, this game can serve as a framework for experimental manipulations of social signals. Two types of social signals we will investigate in this study are the degree to which both the initial judgments of a partner and previous experience with that partner can alter decisions of trust and reciprocity.

Research has demonstrated that trustworthiness is often rapidly inferred from social signals, and can in turn influence behavior in the Trust Game. Trustworthiness judgments are influenced by brief social interactions (Frank, Gilovich, & Regan, 1993), as well as information about an individual's moral character (Delgado, Frank, & Phelps, 2005). Even more subtly, signals of trustworthiness can be detected from simply viewing faces (Winston, Strange, O'Doherty, & Dolan, 2002). Facial expressions can be processed outside of conscious awareness (Morris et al., 1998), and indeed, competence judgments about an individual can be made within 100 ms (Willis & Todorov, 2006) and affective judgments about an individual can be made as quickly as 140 ms (Pizzagalli et al., 2002). Individuals who are attractive or who appear happy are also more likely to be viewed as trustworthy (Scharlemann, Eckel, Kacelnik, & Wilson, 2001). Our group has recently investigated how initial

Please cite this article in press as: Chang, L. J., et al. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology* (2010), doi:10.1016/j.cogpsych.2010.03.001

impressions can influence trust, and demonstrated that implicit judgments of facial trustworthiness can predict the amount of financial risk a person is willing to take in a Trust Game (van 't Wout & Sanfey, 2008). In this study, normed ratings of player trustworthiness (as assessed by briefly viewing a photograph of each player) were a significant predictor of how much money these players were given in a standard one-shot Trust Game. These set of studies support the notion that both explicit (e.g., information about a partner's moral character) and implicit social signals (e.g., facial trustworthiness) can influence initial judgments of trustworthiness, and that these judgments can in turn impact the degree to which people actually place trust in other individuals in a meaningful social interaction.

Social signals can also be inferred from repeated interactions. The best predictor of whether a person will place trust in their partner in a given Trust Game round is whether or not this partner previously reciprocated trust (King-Casas et al., 2005). The process of placing trust when it has previously been reciprocated, but stopping once trust is abused, is often referred to as a tit-for-tat strategy, and has been demonstrated to be the optimal strategy for repeated interactions (Axelrod & Hamilton, 1981). Repeated interactions have also been shown to influence subjective ratings of moral character in a Prisoner's Dilemma game (Singer, Kiebel, Winston, Dolan, & Frith, 2004) and in a Trust Game (Delgado et al., 2005). These findings suggest that in a repeated interaction, trustworthiness can be learned based on the history of a partner's behavior.

One model of investor behavior in the context of a Trust Game therefore involves an initial judgment of trustworthiness based on available information, which is then updated based on subsequent interactions with that partner. One question that currently remains unanswered is the interaction between this initial assessment and the subsequent updating, for example, the degree to which each signal may contribute to the final trust decision, and how concordant (a trustworthy face engaged in reciprocal behavior) and conflicting (e.g., a trustworthy face who does not reciprocate) information is handled. No study has directly investigated this question in the context of a Trust Game, though one experiment has provided preliminary evidence suggesting that initial judgments may influence the way information from repeated interactions is updated (Delgado et al., 2005). In this study, participants played a repeated Trust Game with three fictional characters. Prior to interacting with these purported partners, participants were given a short vignette describing the moral character of each partner. One character was depicted as "good", one "neutral", and a third as "bad". The investigators observed that participants rated the "good" character as more trustworthy at the start of the game, and were in turn more likely to trust them. However, because all partners reciprocated 50% of the time, participants learned to trust the "good" partners less over time, and in fact began to "match" the 50% reinforcement probability (Herrnstein, 1961). At the conclusion of the game, even though participants trusted the "good" partner less than they did at the beginning, they still placed more trust in them than either of the other partners, and were still investing more than 60% of the time. There are two possible interpretations of this finding from a reinforcement learning framework. First, as suggested by the authors, the positive moral information may have biased the participants to ignore negative feedback, meaning they were unable to update the value of the partner after they were betrayed. An alternative interpretation is that the positive moral information increased the initial trust evaluation of the partner, but did not influence the way the participant interpreted the feedback. According to this interpretation, if given enough trials, the participant would have eventually learned the 50% reinforcement rate, though it would have taken longer compared to the neutral and negative partners. However, the design employed in this study makes it difficult to assess which hypothesis is more likely.

A useful method to examine the question of how initial judgment and experience interact is to employ mathematical models of behavior. Reinforcement learning (RL) is concerned with understanding how people learn from feedback in repeated interactions with the environment (Sutton & Barto, 1998). Assuming that the decision-maker is attempting to maximize his or her reward on each trial, one strategy is to predict the value of an environmental state, and then update these predictions based on the actual feedback received. One method for updating predicted values is to use the simple Rescorla–Wagner delta rule (Rescorla & Wagner, 1972), which quantifies on each trial the difference between the predicted value V and actual reward r, with this difference referred to as prediction error.

$$\delta = r_S - V_S(t) \tag{1}$$

Please cite this article in press as: Chang, L. J., et al. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology* (2010), doi:10.1016/j.cogpsych.2010.03.001

The most straightforward way to learn the value of the relevant stimulus s, is to update its predicted value in proportion to the current prediction error δ. The degree to which the prediction error influences the new value is scaled by a learning rate α, where $0 < \alpha < 1$.

$$V_S(t+1) = V_S(t) + \alpha\delta \tag{2}$$

Thus, receiving rewards greater than expected will lead one to increase the value associated with a given stimulus. Conversely, receiving rewards that were less than expected will cause a decrease in that value. Using a RL approach, all stimuli have an initial starting reward value, which is updated via a learning rule. Because of its simplicity, this framework not only provides a very powerful way to understand how people learn from feedback, but also provides a principled way to understand how social signals influence learning in a repeated Trust Game.

Use of this framework to understand how people learn in a social context encourages very specific hypothesis testing, and has the potential to provide insight into the subtle processes involved in social learning. To date, relatively few studies have attempted to study social learning from an RL perspective (Behrens, Hunt, Woolrich, & Rushworth, 2008; King-Casas et al., 2005). However, some recent studies have begun to use modeling in conjunction with behavior to better understand how social decision-making develops. For example, one experiment (Hampton, Bossaerts, & O'Doherty, 2008) used computational modeling to provide insight into the process of mentalizing about another player's strategy in a game known as the Inspection Game. Additionally, Apesteguia, Huck, and Oechssler (2007) demonstrated that when given the opportunity to view other player's behavior in a game, people will often imitate the strategy that provides the highest payoff. Greater discrepancies between an individual's payoff and another player's payoff result in an increased likelihood of switching to the other strategy.

Finally, a few recent studies have utilized computational approaches to study how social advice can impact learning (Biele, Rieskamp, & Gonzalez, 2009; Doll, Jacobs, Sanfey, & Frank, 2009). In these studies, prior to a standard learning task, participants are given information (termed as advice or instructions) from either another participant or from the experimenter about the optimal choice. These experiments have found evidence supporting the notion that social information leads to learning biases, namely that accurate information helps participants learn better, while inaccurate information impairs learning. Biele et al. (2009) found support for a model that assigned greater weight to outcomes consistent with the advice than to the same outcomes on unadvised choices. Doll et al. (2009) found that the best fit of the behavioral data was produced by a model that initialized instructed stimuli to a higher than normal starting value and reduced the impact of instruction inconsistent outcomes while increasing the impact of instruction consistent outcomes. These studies suggest that *explicit* information such as advice or moral information can impact not only initial expectations but also how people learn from feedback. Information consistent with the prior information is weighted higher in the value update, and information inconsistent with the advice is weighted lower. However, no study to date has examined how *implicit* information impacts learning in an interactive social decision scenario.

The present study adapted the design of van 't Wout and Sanfey (2008) to examine how implicit initial trustworthiness information (i.e. facial features) interacts with experienced trustworthiness (i.e. the probability of reciprocation) in a repeated Trust Game. First, we expected to replicate our previous finding that facial trustworthiness influences initial financial risk-taking in a social context (van 't Wout & Sanfey, 2008). Second, we expected to replicate other work, which has demonstrated that previous experiences also influence behavior (Axelrod & Hamilton, 1981; King-Casas et al., 2005). Finally, and most importantly, we predicted that these two processes, facial trustworthiness and experienced trustworthiness, would interact such that partners that both look trustworthy and reciprocate frequently will be entrusted with the most money. To increase our construct validity, we employed multiple measurements of trustworthiness, which included behavior in the Trust Game as well as subjective ratings. To further characterize our behavioral findings, we used RL models to test three distinct processing hypotheses – (1) initialization, (2) confirmation bias, and (3) dynamic belief. The Initialization models (GL initialization & trust decay) posit that the implicit trustworthiness judgments influence behavior at the beginning of the game, but are eventually overridden by the player's actual experiences (i.e. whether or not trust is reciprocated). The Confirmation Bias model proposes

Please cite this article in press as: Chang, L. J., et al. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology* (2010), doi:10.1016/j.cogpsych.2010.03.001

that initial implicit trustworthiness judgments influence the way feedback (i.e. non-reciprocated trust) is updated throughout the interactions (Biele et al., 2009; Delgado et al., 2005; Doll et al., 2009). This model assumes that learning is biased in the direction of the initial impressions. Finally, the Dynamic Belief model proposes that the facial trustworthiness judgment serves as an initial trustworthiness belief, which is continuously updated based on the player's experience in the game. These beliefs, in turn, influence learning. This model equally emphasizes the initial judgment and experience and predicts that players will learn to give more money to partners that are trustworthy, and less money to partners that betray trust. By explicitly formalizing the potential mechanisms via these models, this study can increase our understanding of how trust is placed in social economic exchanges.

2. Methods

2.1. Participants

Sixty-four undergraduates were recruited from the psychology participant pool at the University of Arizona and received course credit for their participation in the experiment. Three participants were excluded after indicating during debriefing that they did not understand the experiment, leaving a total of 61 participants (mean age = 18.67 sd = 1.38, female = 79%). All participants gave informed consent, and the study was approved by the local Institutional Review Board.

2.2. Trust Game

Participants played a repeated Trust Game in the role of Player A as described in the introduction. We employed a 2×2 within-subjects design in which partner's level of facial trustworthiness (high or low) was crossed with partner's level of experienced trustworthiness (high or low). Each Player B represented one of the experimental conditions. Level of facial trustworthiness was assessed via independent ratings of partner photographs (see below). We defined level of experienced trustworthiness as a high (80%) or low (20%) probability of reciprocating an offer. Money invested by the participant was multiplied by a factor of 4. If an offer was reciprocated, Player B always reciprocated 50% of the total multiplied amount sent by Player A. When trust was not reciprocated, Player B did not return any of the multiplied amount of money back to Player A. Participants played 15 randomly ordered interspersed rounds with each partner (60 trials total, plus 30 slot machine gambles). On each trial, participants were endowed with $10. Each trial lasted 16 s and began with a short fixation cross (1000 ms) followed by a picture of the partner (3500 ms). Participants then decided how much money they wanted to invest by scrolling through offers that randomly increased or decreased in $1 increments. After submitting their investment, participants were shown the amount of money they chose to invest (multiplied by 4). Player B's decision to either keep all of the money or reciprocate was then revealed along with a summary of the payoffs to each player. If the participant did not submit an offer in time (8000 ms), they forfeited all of their money for the round. As a non-social control, participants also had an opportunity to "gamble" with two different slot machines. Just like the Trust Game trials, participants were allowed to invest any amount of their $10 endowment in $1 increments. One slot machine paid out twice the investment with an 80% probability and the other slot machine paid out at a 20% probability. Thus, the slot machine trials were identical to the Trust Game trials except that they offered a non-social context.

2.3. Stimuli

The stimuli for the human partners were selected from the Winston stimuli set (Adolphs, Tranel, & Damasio, 1998) based on trustworthiness ratings from a previous study (van 't Wout & Sanfey, 2008). Two sets of four faces were selected that were matched for trustworthiness and attractiveness. Each set consisted of one male and one female that were previously rated high on trustworthiness (mean = 4.10) and low on trustworthiness (mean = 3.10) on a 7-point Likert scale.

Please cite this article in press as: Chang, L. J., et al. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology* (2010), doi:10.1016/j.cogpsych.2010.03.001

Participants were randomly assigned to play with one of the picture sets, with the other set serving as a control. For each participant the high and low trustworthy pictures were randomly assigned to an experienced trustworthiness condition (high vs. low). The four cells were balanced across subjects, χ^2 (3) = 0.56, p = 0.91. This ensured that any observed effects could not be attributable to a single picture. At the end of the game participants rated both sets of stimuli – those they played against and those they did not – on trustworthiness, attractiveness, competence, aggressiveness, and likeability using a 5-point Likert scale. Participants were also asked to estimate the percentage of the time that each of the players they played with reciprocated their offers, from 0% to 100% in 10% increments. All stimuli were presented on a laptop via EPrime software (Psychology Software Tools, Inc., Pittsburgh, PA).

3. Results

3.1. Behavioral data

Overall, as expected, we found a main effect of reciprocity, where participants gave more money overall to partners who reciprocated 80% of the time (mean = 5.64, se = 0.18) as compared to partners who only reciprocated 20% of the time (mean = 3.42, se = 0.20) using a repeated measures ANOVA $F(1, 60) = 125.70$, $p < 0.001$, $\eta^2 = 0.68$. There was a trend for partner type that approached significance, $F(2, 120) = 2.41$, $p = 0.09$ where participants tended to invest more money in participants who looked more trustworthy (mean = 4.76, se = 0.17) as compared to both those who looked untrustworthy (mean = 4.44, se = 0.17) and computer controls (mean = 4.38, se = 0.23). There was a significant partner by reciprocity interaction $F(2, 120) = 5.34$, $p = 0.006$, $\eta^2 = 0.08$, such that participants invested the most money in high trustworthy partners that reciprocated 80% of the time (mean = 6.11, se = 0.23) and the least amount of money in low trustworthy partners that reciprocated 20% of the time (mean = 3.26, se = 0.22) (see Fig. 1).

Additionally, we successfully replicated our previous finding (van 't Wout & Sanfey, 2008), in which facial trustworthiness impacted the amount of money invested on the first trial of each pairing, $F(2, 120) = 3.22$, $p = 0.04$, $\eta^2 = 0.05$. In their first interaction, participants invested significantly more money in high trustworthy looking partners (mean = 5.1, se = 0.23) as compared to low trustworthy looking partners (mean = 4.37, se = 0.23), $p = 0.03$ (see Fig. 2, Panel A).

Finally, the interaction that we observed in our main behavioral results was not completely driven by a strong effect of facial trustworthiness at the beginning of the experiment. We observed a significant interaction between facial trustworthiness and probability of reciprocation on the last trial of the experiment $F(2, 120) = 4.13$, $p = 0.02$, $\eta^2 = 0.06$, suggesting that the effect persists throughout all 15 trials (see Fig. 2, Panel B). These data were log transformed to account for negative skew in the data.

3.2. Post-experiment ratings

As a manipulation check to ensure that the pictures used in each condition were viewed appropriately, we compared all of the ratings for each participant's control picture set. These picture sets were counterbalanced across participants, so that control sets served as the experimental pictures for other participants. Mixed effects regression revealed that participants rated those partners that were selected to look more trustworthy as actually being more trustworthy $b = 0.44$ (se = 0.17), $t = 2.66$, $p < 0.05$, more attractive $b = 0.60$ (se = 0.11), $t = 5.25$, $p < 0.05$, more likeable $b = 0.56$ (se = 0.14), $t = 3.90$, $p < 0.05$, more competent $b = 0.39$ (se = 0.12), $t = 3.12$, $p < 0.05$, and less aggressive $b = -0.83$ (se = 0.13), $t = -6.20$, $p < 0.05$ then those that were selected to look untrustworthy.

To determine the effect of the repeated interaction on post-game trustworthiness judgments, we examined the main effects of the initial trustworthiness judgment, the probability of reciprocation, and their interaction. All predictions were supported (see Fig. 3). There was a significant main effect of partner type, $F(2, 120) = 7.05$, $p < 0.001$, $\eta^2 = 0.10$. Partners that were selected to look more trustworthy based on ratings from an independent sample (mean = 3.10, se = 0.09) were rated as more trustworthy as compared to partners that were selected to look untrustworthy (mean = 2.66,

Please cite this article in press as: Chang, L. J., et al. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology* (2010), doi:10.1016/j.cogpsych.2010.03.001

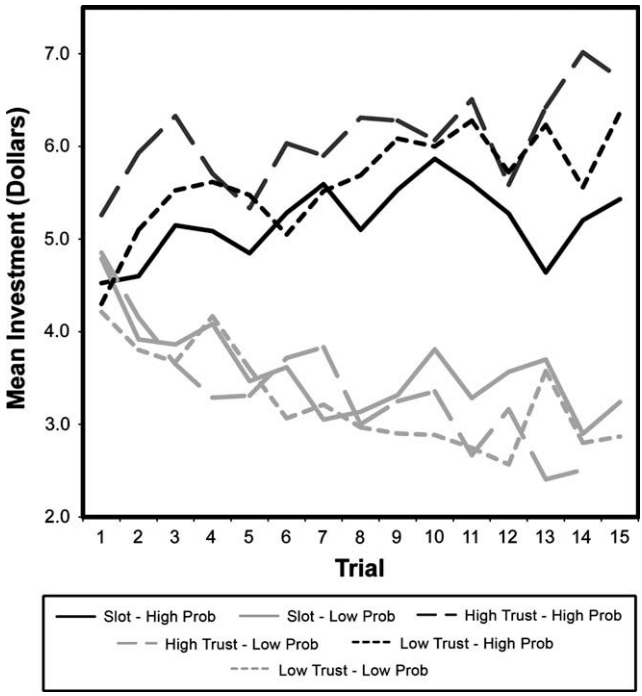

Fig. 1. Overall learning in the iterated Trust Game. Note: this figure shows the mean investment amount across subjects for each trial of each condition.

se = 0.09), $p < 0.05$. Partners that reciprocated more frequently (mean = 3.61, se = 0.07) were rated as more trustworthy then those that reciprocated infrequently (mean = 2.18, se = 0.08), $F(1, 60) = 183.16$, $p < 0.001$, $\eta^2 = 0.75$. Finally, there was a significant interaction between initial trustworthiness and probability of reciprocation, $F(2, 120) = 8.58$, $p < 0.001$, $\eta^2 = 0.13$, such that partners that both looked trustworthy and reciprocated frequently were rated the most trustworthy (mean = 4.02, se = 0.11) and partners that both looked untrustworthy and reciprocated infrequently were rated the least trustworthy (mean = 1.93, se = 0.12). Also, it is interesting to note that partners that looked trustworthy and reciprocated frequently were rated as more trustworthy (mean = 4.02, se = 0.11) in comparison to both novel faces matched for trustworthiness (mean = 3.43, se = 0.08) $t(60) = 4.20$, $p < 0.001$, and for untrustworthy partners that reciprocated at the same frequency (mean = 3.38, se = 0.14), $t(60) = 3.56$, $p < 0.001$. Finally, there was no significant difference in trustworthiness ratings between untrustworthy partners that reciprocated frequently (mean = 3.38, se = 0.14) as compared to trustworthy control pictures (mean = 3.43, se = 0.08), $t(60) = -0.35$, $p > 0.05$, suggesting that positive experiences can override initial negative impressions.

Participants were quite accurate in their estimation of their partners' behavior, as gauged by the percentage of the time they believed the partners sent money back. Participants estimated that partners in the high probability condition (mean = 65.19, se = 1.55) reciprocated more often then partners in the low probability condition (mean = 28.08, se = 1.74) $F(1, 60) = 186.09$, $p < 0.001$, $\eta^2 = 0.76$. Participants did not differ in their probability estimation as a function of their partner's trustworthiness $F(2, 120) = 1.0$, $p > 0.05$. However, there was a significant partner by probability interaction $F(2, 120) = 3.47$, $p = 0.03$, $\eta^2 = 0.06$, where high trustworthy looking partners (mean = 70.82, se = 2.3) were estimated to reciprocate more frequently than both low trustworthy partners (mean = 62.95, se = 2.48) $t(60) = 2.29$, $p = 0.03$ and computer controls (mean = 61.80, se = 3.03) $t(60) = 2.57$, $p = 0.01$. This indicates that participants explicitly judged the probability of reciprocation to be higher for partners that looked more trustworthy in comparison to other partners.

Please cite this article in press as: Chang, L. J., et al. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology* (2010), doi:10.1016/j.cogpsych.2010.03.001

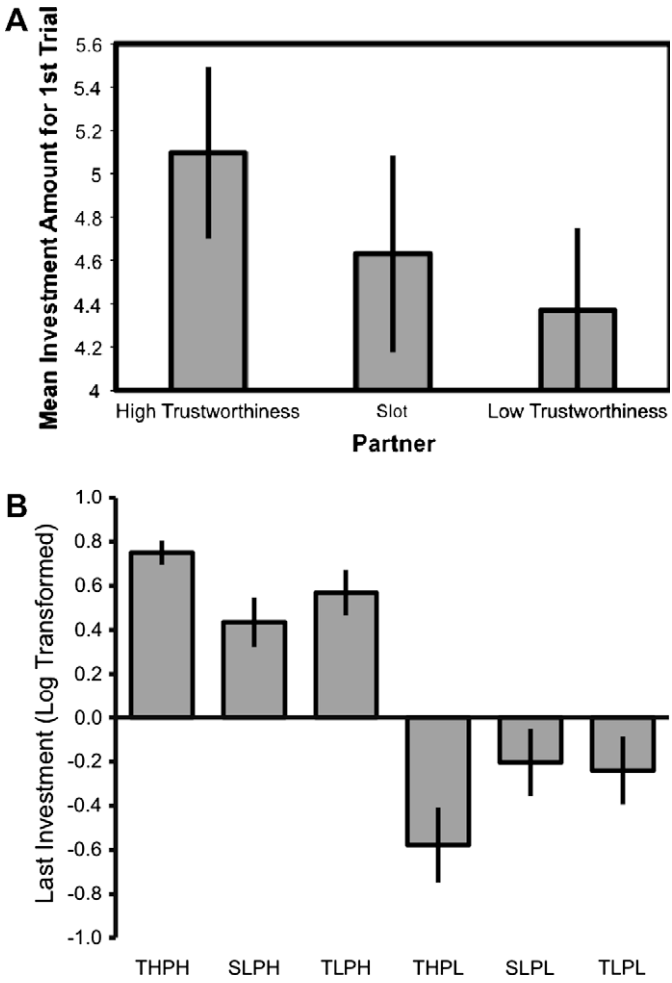

Fig. 2. First and last encounters mean offer amount. Note: (A) mean investment for first trial by partner. (B) Mean investment for last trial for each condition. Data was log transformed due to extreme negative skew. THPH = high trustworthiness and high probability. THPL = high trustworthiness and low probability. TLPH = low trustworthiness and high probability. TLPL = low trustworthiness and low probability. SLPH = slot machine and high probability. SLPL = slot machine and low probability. Error bars represent ±1 standard error.

3.3. RL models

Our results demonstrate the notable effect that initial perceptions of trustworthiness interact with experience to influence both the amount of trust actually placed in a partner and the perceived judgments of trustworthiness revealed via participants' post-experiment subjective ratings. However, these analyses cannot speak to *how* these two variables might be interacting. There are several plausible explanations for this effect. First, initial trustworthiness judgments might merely influence the starting trust value, but not how participants update their beliefs after feedback. Second, it is possible that the initial expectations influence how feedback is interpreted, where feedback that is consistent with the expectation (e.g., cooperation by partners that are perceived to be trustworthy) is weighted more heavily than feedback that is inconsistent with the expectation (e.g., cooperation by partners that are perceived to be untrustworthy). Third, it is possible that initial expectations influence how

Please cite this article in press as: Chang, L. J., et al. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology* (2010), doi:10.1016/j.cogpsych.2010.03.001

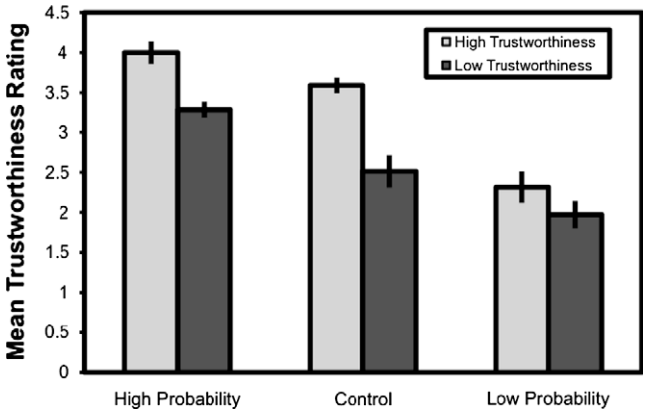

Fig. 3. Post-experiment subjective trustworthiness ratings. Note: error bars represent ±1 standard error.

feedback is interpreted, but the feedback can also in turn influence the expectations. For example, partners that repeatedly violate trust are eventually perceived to be untrustworthy, which means that feedback suggesting otherwise (i.e. cooperation) will be less likely to change participants' behavior. In an effort to further characterize our behavioral results, we employed a class of RL models referred to as value learning (Sutton & Barto, 1998) to test these different computational accounts. These models attempt to calculate how an individual learns the value V of a stimulus S for a given trial t using a minimal number of parameters. In our approach, value refers not to the probability of making a discrete decision, but to the actual value of a given partner's trustworthiness (in dollars). We utilized a cross-validation procedure that is robust to differences in model complexity to compare three different processing hypotheses (initialization, confirmation bias, and dynamic belief) against a baseline model.

3.4. Gain Loss model (baseline)

Considerable evidence supporting Prospect Theory (Kahneman & Tversky, 1979) has demonstrated that people prefer avoiding losses as compared to acquiring gains of the same magnitude. Given these findings, we use a model that differentially updates gains and losses via separate learning rates as a baseline model (Doll et al., 2009; Frank, Moustafa, Haughey, Curran, & Hutchison, 2007; Yechiam, Busemeyer, Stout, & Bechara, 2005). This model computes a predicted value for the next trial for each stimulus based on the experienced outcome:

$$V_s(t+1) = V_s(t) + \alpha_G \delta^+ + \alpha_L \delta^- \tag{3}$$

where α_G is the amount that a positive outcome (notated by the + superscript) is weighted and α_L is the amount that a negative outcome (notated by the − superscript) is weighted in the update ($0 < \alpha < 1$). This allows people to learn from losses differently than gains, which we think is particularly important due to the social nature of the task. For example, if a participant believes their goodwill has been violated, they will likely adapt their behavior quickly (Bohnet & Zeckhauser, 2004). We chose to set the initial value $V_s(1)$ for all conditions to the average amount sent by the participants on the first trial of the game (mean = 14.52). This value is calculated by inputting an initial investment of \$4.52 into Eq. (9).

3.5. Gain Loss Initialization model

We used two different models to test the initialization hypothesis (see Trust Decay model below for alternative approach). This model initialized the starting values of the baseline GL model based on the

Please cite this article in press as: Chang, L. J., et al. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology* (2010), doi:10.1016/j.cogpsych.2010.03.001

trustworthiness ratings from an independent sample. The values for high trustworthiness (T_{HT} = 3.6) and low trustworthiness (T_{LT} = 2.5) were scaled by a free parameter λ when t = 1, where $0 < \lambda < 20$.

$$V_S(t) = \begin{cases} 14.52 + \lambda \cdot T_{HT} \\ 14.52 - \lambda \cdot T_{LT} \end{cases} \tag{4}$$

This model predicts that the perceived facial trustworthiness will only influence the initial expectations and will have no bearing on the update rule.

3.6. Confirmation Bias model

Previous research has examined the effect of explicit information on decision-making (Biele et al., 2009; Doll et al., 2009). These studies have proposed models that give a higher weight to feedback that is consistent with the advice, and lower to feedback inconsistent with the advice. To examine this hypothesis we will test a model that is similar to Doll et al.'s (2009) instructed learning model and Biele et al.'s (2009) outcome bonus model, formally defined as

$$V_S(t+1) = V_S(t) + \alpha_G \delta^+ + \alpha_L \delta^- + \phi[T_S]^{\text{repay}} - \phi[1 - T_S]^{\text{abuse}} \tag{5}$$

where $0 < \phi < 10$. In this model, participants receive a fixed bonus that is proportional to their partner's level of facial trustworthiness T_S scaled by a free parameter ϕ, when their partner reciprocates their investment (denoted by the 'repay' superscript). If their partner fails to reciprocate, this amount serves as a deduction (denoted by the 'abuse' superscript). This means that higher levels of facial trustworthiness will promote quicker learning from gains, while lower facial trustworthiness will facilitate greater learning from losses. Facial trustworthiness was determined by ratings from an independent sample that were transformed to range between 0 and 1 (T_{HT} = 0.72 for trustworthy faces; T_{LT} = 0.5 for untrustworthy faces).[1] This model tests the confirmation bias hypothesis, and predicts that initial facial trustworthiness judgments will influence how feedback is interpreted consistently over the course of the experiment.

3.7. Trust Decay model (initialization)

An alternative hypothesis is that facial trustworthiness judgments influence initial behavior, but then become less important with increased experience with a partner. This is a different test of the initialization hypothesis from the Gain Loss Initialization model because it predicts that facial trustworthiness will influence the update early on and not just start at a higher value. This model decreases the influence of the trustworthiness bonus as a function of time by ρ and is formally defined as

$$V_S(t+1) = V_S(t) + \alpha_G \delta^+ - \alpha_L \delta^- + e^{-\rho \cdot t}[T_S]^{\text{repay}} - e^{-\rho \cdot t}[1 - T_S]^{\text{abuse}} \tag{6}$$

where $0 < \rho < 1$. This model is similar to the Confirmation Bias model initially, but exponentially decays the influence of the trustworthiness bonuses and deductions over time. This model tests the initialization hypothesis (also see Gain Loss Initialization model) and predicts that facial trustworthiness judgments will provide a preliminary estimate of an individual's level of trustworthiness, but will eventually be overcome by experience – a prediction that has previously been framed from a multiple systems perspective (Frank et al., 2007).

3.8. Dynamic Belief model

The final hypothesis that we will test also treats trustworthiness as a bonus in the update function, but rather than being a fixed bonus based on the initial trustworthiness judgment like the Confirmation Bias and Trust Decay models, it adapts over time based on the perceived level of trustworthiness.

[1] This model uses an additive bonus, as was employed by Biele et al. (2009). We were unable to get a multiplicative implementation of Doll et al.'s (2009) Instructed Learning Model to converge.

Please cite this article in press as: Chang, L. J., et al. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology* (2010), doi:10.1016/j.cogpsych.2010.03.001

This means that the model can learn the level of trustworthiness T for each partner S and will use this information as a bonus or deduction in the update. This model is formally defined as

$$T_S(t+1) = T_S(t) + \phi[\text{Outcome}(t) - T_S(t)] \tag{7}$$

$$V_S(t+1) = V_S(t) + \alpha_G\delta^+ + \alpha_L\delta^- + \theta\left[T_S(t+1)^{\text{repay}} - T_S(t+1)^{\text{abuse}}\right] \tag{8}$$

where ϕ represents the trustworthiness learning rate, and θ is a free parameter used to scale the amount of influence of the trustworthiness bonus in the value update and Outcome = 1 if partner reciprocates trust or 0 if the partner abuses trust. This model can dynamically learn the trustworthiness of each partner and will add a bonus that is proportional to the level of perceived trustworthiness if the partner reciprocates, or alternatively will deduct a value proportional to the level of trustworthiness if the partner defects. This model differs from the Confirmation Bias model because defections by a partner with lower perceived trustworthiness result in smaller deductions. Another important conceptual distinction is that the model allows facial trustworthiness to influence the update like the other two bonus models, but it also allows feedback to influence the trustworthiness beliefs. Because this model learns the perceived trustworthiness of each partner, it can be used to predict each individual participants' subjective trustworthiness ratings that were measured at the end of the experiment. This model tests a processing hypothesis and predicts that perceived trustworthiness changes over time but is continually used to update the value associated with a given partner.

3.9. Model evaluation

To evaluate the models, we employed a cross-validation procedure in which the models were initially trained on half of the data and then subsequently tested on the other half. This approach substantially reduces over-fitting and also provides a useful way to compare models of differing complexity. During the training phase, models were fit to the participant's actual behavioral data by minimizing the sum of the squared error (SSE) and parameters were estimated for the entire group. Models were compared using a metric that rewards the most parsimonious model. We then compared the models in their ability to predict the behavioral data out-of-sample using the parameters estimated during the training phase. This approach controls for models having different numbers of free parameters when fitting to behavioral data (Hampton & O'Doherty, 2007).

We calculated the reward for a given stimulus s for trial t using the following equation

$$r_s(t) = \left(10 - \text{Investment} + \frac{4 * (\text{Investment})}{2}\right) \tag{9}$$

Investments were multiplied by a factor of 4 and were divided by 2 because they were split equally between both parties (if they were reciprocated). The participant's reward takes into account the amount of money they kept plus the amount that was reciprocated, which allows the possibility for participants to still receive reward when there is negative prediction error (i.e. the partner did not reciprocate).

Parameters were estimated during the training phase by minimizing the SSE between the behavioral data and the predictions from the various models on every odd trial using fmincon (Coleman & Li, 1996), a multivariate constrained nonlinear optimization algorithm implemented in Matlab (Mathworks, Cambridge, MA).

$$\sum (r_s(t) - V_s(t))^2 \tag{10}$$

The value V for a given state s at time $t + 1$ is updated using the functions specified above. Because of the small number of trials for each condition ($n = 15$), the parameters were estimated for the entire group. This means that while the models were fit to each participant's individual behavioral data, the error in the parameter estimation was pooled across subjects. This procedure has been previously utilized when individual parameter estimates are not stable as a result of a small number of trials and collinearity between parameters (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006). While we observe

Please cite this article in press as: Chang, L. J., et al. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology* (2010), doi:10.1016/j.cogpsych.2010.03.001

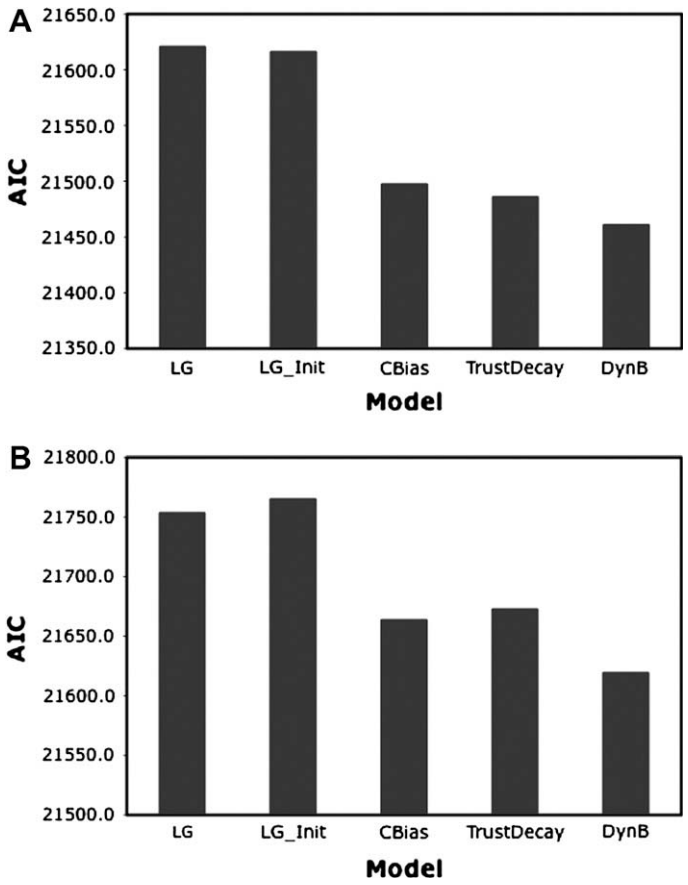

Fig. 4. RL model cross validation. Note: (A) this graph depicts the AIC fits from the training phase, in which each of the models were fit to the odd trials and parameters were estimated for the entire group. LG = baseline Gain Loss model (Eq. (3)) with two parameters, LG_Init = Gain Loss Initialization model (Eq. (4)) with three free parameters, CBias = confirmation bias model (Eq. (5)) with three parameters. TrustDecay = Trust Decay model (Eq. (6)) with three parameters. DynB = Dynamic Belief model (Eqs. (7) and (8)) has four parameters. (B) This graph depicts the AIC fits of the test phase in which the parameters estimated from the odd trials were used to predict the even trials. This out-of-sample test controls for the number of free parameters when comparing model fits because no parameters are actually estimated during this procedure.

a similar hierarchy of model fits when an individual parameter is fit for each participant, we report the more stable group fits.[2] Multiple start locations were used to minimize the risk of the optimization routine getting stuck in local minima.

All models were compared to a baseline RL model (Eq. (3)), using the Akaike Information Criteria (AIC), a value that provides a metric of model fit by taking into account the complexity of the model (i.e. the number of estimated parameters).[3] AIC rewards the most parsimonious model by penalizing for additional free parameters and is formally defined as

[2] It is important to note that our models are accounting for each individual participant's trial-by-trial sequence of choices and outcomes despite estimating the model parameters for the group. As a demonstration, randomly shuffling the trial sequence for each participant using the Gain Loss Baseline model results in a dramatic decrease in model fit (Normal Sequence AIC = 25326.27; Random Sequence AIC = 26600.49).

[3] We also calculated the Bayesian Information Criterion (Schwarz, 1978), which provides a slightly larger penalty for the number of free parameters and observed results consistent with the AIC (LG = 6859.94; LG Initialization = 6871.57; Confirmation Bias = 6770.03; Trust Decay = 6779.08; Dynamic Belief = 6725.86).

Please cite this article in press as: Chang, L. J., et al. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology* (2010), doi:10.1016/j.cogpsych.2010.03.001

$$AIC = 2k + n\left[\ln\left(\frac{2\pi RSS}{n}\right) + 1\right] \quad (11)$$

where k is the number of free parameters, n is the number of observations, and RSS is the residual sum of squares (Akaike, 1974).

To provide a more stringent model comparison procedure, we then tested the models out-of-sample by using the parameters estimated from the training phase and minimizing the SSE between the behavioral data and the model predictions on the even trials. Because no parameters are actually being estimated during this process, the model fits cannot be artificially inflated as a result of additional free parameters. While this type of procedure is often used in financial forecasting to predict the last 20% of the data, we chose to evenly sample throughout the trials to avoid unfairly penalizing our models that predict changes in the update function as a function of experience (e.g., trust decay and dynamic belief).

3.10. Modeling results

Overall, most of our models provided a better explanation of the data than the baseline GL model (AIC = 21620.29; RSS = 18895.58, N = 5248; see Fig. 4, Panel A). Both the Confirmation Bias model (AIC = 21496.84; RSS = 18449.23, N = 5248) and the Trust Decay model (AIC = 21485.47; RSS = 18409.31, N = 5248) provided a better fit of the data than the baseline GL model and the Dynamic Belief model (AIC = 21460.57; RSS = 18315.19, N = 5248) provided the best fit of all models tested. The one exception was the GL Initialization model, which did not appear to fit the data any better than the baseline model (AIC = 21615.64; RSS = 18871.64, N = 5248). These results suggest that facial trustworthiness judgments do not just merely influence the initial value, but rather seem to affect how feedback is interpreted. The Dynamic Belief model, which allows the initial expectation to influence the update function and also allows for the feedback to update the expectation, appeared to be the best account of the behavioral data. However, it is important to note that all of these models have a different number of free parameters, with the Dynamic Belief model having the most. While the AIC and BIC metrics are standard ways of penalizing for additional free parameters, a stronger test is to look out-of sample.

We find a similar hierarchy of results in our out-of-sample prediction procedure (see Fig. 4, Panel B). The GL Initialization model (AIC = 21764.75; RSS = 19437.74, N = 5248) fits about the same, in fact slightly worse, than the baseline GL model (AIC = 21753.11; RSS = 19394.69, N = 5248). The Confirmation model (AIC = 21663.21; RSS = 19065.27, N = 5248) and the Trust Decay model (AIC = 21672.25; RSS = 19098.15, N = 5248) both fit better than the baseline model and the Dynamic Belief model (AIC = 21619.04; RSS = 18905.48, N = 5248) exhibits the best fit. These results provide additional evidence that initial beliefs about trustworthiness influence the update function, and that feedback in turn can update the trustworthiness beliefs. A simulation of the Dynamic Belief model using the parameters estimated from the training phase can be seen in Fig. 5.

Consistent with our prediction, participants appeared to adapt their behavior more radically when their partner defected (mean α_L across models = 0.35) as compared to when their partner cooperated (mean α_G across models = 0.02). In addition, there seems to be little evidence that high initial values can explain our behavioral results. The Gain Loss Initialization model which specifically manipulated the starting value based on the trustworthiness ratings did not explain the data any better than the baseline model. Also, the Trust Decay model, which only uses the trustworthiness bonuses early on, did not explain the data any better than the Confirmation Bias model. In fact, the parameter that was estimated for the decay, ρ, was essentially 0 indicating that the model reduced to the Confirmation Bias model with a ϕ of 1. A summary of the estimated parameters for each model can be seen in Table 1.

3.11. Predicted trustworthiness ratings

While the Dynamic Belief model appeared to provide the best explanation of the behavioral data, it is also possible to examine how well it can capture subjective perceptions of trustworthiness. The

Please cite this article in press as: Chang, L. J., et al. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology* (2010), doi:10.1016/j.cogpsych.2010.03.001

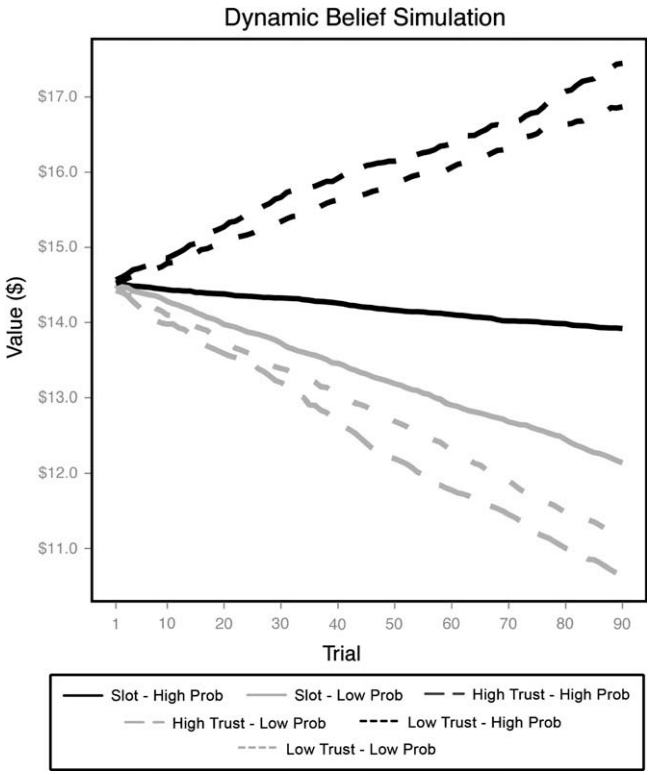

Fig. 5. Dynamic Belief simulation. Note: this figure shows a simulation of the Dynamic Belief model for 60 subjects using the best fitting parameters estimated from the model training procedure. The initial value was the mean initial investment ($4.52) and initial trustworthiness values were the normative trust ratings (high = 0.72; low = 0.5). Outcomes were randomly generated based on the high (80% reciprocate) and low (20% reciprocate) probability structure of the game.

Table 1
Estimated parameters for RL models.

Model	α_G	α_L	λ	ϕ	ρ	θ
LG	0.02	0.37				
LG initialization	0.02	0.37	0.24			
Confirmation bias	0.02	0.33		0.39		
Trust decay	0.02	0.33			0.03	
Dynamic belief	0.01	0.34		0.12		0.45

Note: this table reflects the best fitting parameters from the model training procedure. Because of the limited number of trials used in the estimation procedure ($N = 8$) these parameters should be interpreted with caution.

model makes specific predictions about how trustworthiness beliefs should adapt based on experiences in the game. In this analysis we averaged the last five trials of the Trustworthiness Beliefs (T_S) derived from the learning model and converted them back into ratings (multiplied them by 5) to predict the participants' post-experiment subjective trustworthiness ratings (see Fig. 6). This analysis is useful because it provides a method to test the predictive validity of the Dynamic Belief aspect of the model using participants' actual trustworthiness ratings. To evaluate whether the ratings produced by the model were better than the initial normed trustworthiness ratings at predicting the actual post-task participant ratings, we used a Williams's T2 statistic (Steiger, 1980) to compare

Please cite this article in press as: Chang, L. J., et al. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology* (2010), doi:10.1016/j.cogpsych.2010.03.001

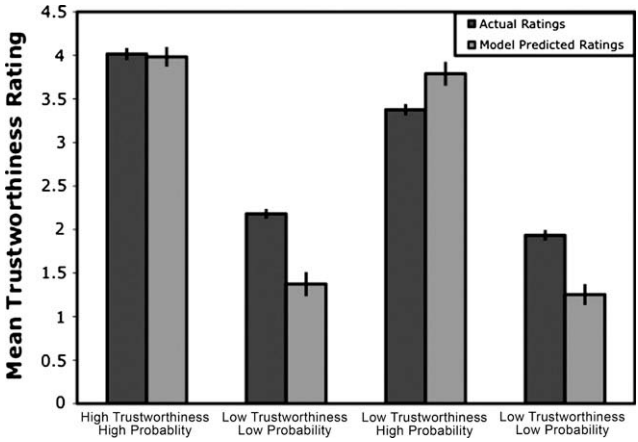

Fig. 6. Model-predicted trustworthiness ratings. Note: this figure shows the trustworthiness ratings predicted by the model (average of last five trials for stability) compared to those reported by the participants. Error bars represent ±1 standard error.

the magnitude of the standardized beta values derived from a mixed effects regression (Baayen, Davidson, & Bates, 2008). The regression was performed on data that was z-transformed and allowed participants' intercepts to vary randomly. Consistent with our hypothesis, this analysis found that the trust ratings predicted from the model $\beta = 0.62$ (se = 0.05) were a better account of the participant's actual post-experiment trust ratings than were the normative trust ratings $\beta = 0.13$ (se = 0.05), $t(57) = 3.15$, $p < 0.05$. This effect appeared to be specific to trustworthiness, as the model derived ratings did not predict other ratings better than the normative trust ratings such as attractiveness $t(57) = 0.47$, ns, aggressiveness $t(57) = 0.94$, ns, competence $t(57) = 0.66$, ns, and likeability $t(57) = 1.85$, ns. In addition, the model predicted the participants' reported probability of reciprocation better than the normed ratings, $t(57) = 3.94$, $p < 0.05$, indicating that the model-predicted ratings captured subjective perceptions of the experienced probabilities. Thus, the Dynamic Belief model appears to not only predict participant's behavior in the game, but it can also predict how their perceptions of trustworthiness change after interacting with a partner.

4. Discussion

This study investigated the processes underlying the decision to trust (or not trust) a partner in a consequential interaction. Previous research has reported that both initial impressions (Delgado et al., 2005; van 't Wout & Sanfey, 2008) and direct experience (King-Casas et al., 2005; Singer et al., 2004) play important roles in influencing judgments of trustworthiness. This experiment provides the first account of how these variables interact in a social interactive financial investment game that has been explicitly designed to study trust (Berg et al., 1995). Consistent with our hypotheses, both the initial trustworthiness judgment of a partner as well as subsequent experience with that partner synergistically influence behavior in this game, in terms of how much money players were willing to entrust to their partner.

4.1. Behavioral measures of trust

Consistent with our group's previous finding, we found that facial trustworthiness influenced participant's initial investment amount (van 't Wout & Sanfey, 2008). On the first round, participants invested more money if their partner looked trustworthy than if the partner looked untrustworthy. This provides further support to the notion that social signals can be conveyed through facial expressions (Oosterhof & Todorov, 2008), and that participants are sensitive to differences in perceived

Please cite this article in press as: Chang, L. J., et al. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology* (2010), doi:10.1016/j.cogpsych.2010.03.001

trustworthiness. In our study, it appears that participants believe trustworthy faces predict a higher probability of reciprocation, and therefore facial trustworthiness may serve as a risk signal which influences the amount of money an individual expects to be sent back. However, there is likely something unique to the social nature of this signal as it is able to be selectively manipulated (compared to pure risk) using a hormone induction (Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005). This hormone, known as oxytocin, acts as a neurotransmitter in the brain and is likely mediating the effect on trust via the amygdala (Baumgartner, Heinrichs, Vonlanthen, Fischbacher, & Fehr, 2008). We also replicated findings which indicate that people use experience as a basis for their trustworthiness judgments (King-Casas et al., 2005). Over time, participants in our study learned to invest more money in partners that reciprocated frequently, and less money in partners that reciprocated infrequently. Together these results suggest that investment behavior in the Trust Game is influenced by both implicit social signals, revealed here through facial trustworthiness, and also by direct social signals conveyed via experience in the game. Finally, these variables appear to act synergistically to influence behavior. Partners that were initially viewed as more trustworthy, and actually turned out to be more trustworthy, were entrusted with the most money in the game, indicating the twin influences of both first impressions and experience.

One potential limitation to our study is that we did not ask whether participants believed they were playing a "real person" at the conclusion of the study. It is unlikely that many participants approached this task believing the person they saw was an active participant due to the rather dated nature of the photographs. However, people do appear to easily anthropomorphize when stimuli behave like real people (Blakemore & Decety, 2001), and more pointedly, the pattern of behavior that we observed is similar to all other reported versions of the repeated Trust Game (Delgado et al., 2005; King-Casas et al., 2005; Krueger et al., 2007). We interpret this to mean that people approached the task at least "as if" they were playing with a real life partner.

4.2. Subjective ratings of trust

Trustworthiness judgments were assessed by participants' behavior towards their partners in the game, and also by their subjective ratings of these partners at the conclusion of the game. Consistent with previous research, we found that post-task subjective trustworthiness ratings were influenced not only by the facial appearance of the partner (van 't Wout & Sanfey, 2008; Winston, Strange, O'Doherty, & Dolan, 2002), but also by the partner's behavior in the game (Delgado et al., 2005; Singer et al., 2004). Partners that reciprocated more frequently were rated as more trustworthy than partners that reciprocated infrequently. We also observed a significant appearance by behavior interaction, where partners that both looked and behaved trustworthily were rated as the most trustworthy of all. These subjective ratings perfectly mirror our behavioral measures of trust.

Several interesting phenomena emerged from examining the post-game ratings of partner trustworthiness. Firstly, partners that looked untrustworthy, but behaved trustworthily, were rated at the same level of trustworthiness as were the control trustworthy faces (with whom participants did not play). Similarly, those partners that were initially viewed as trustworthy, but who behaved in an untrustworthy fashion (i.e. reciprocated trust infrequently), were rated as similar in trustworthiness as the control untrustworthy faces in the post-game subjective ratings (see Fig. 3). This suggests that these fast automatic judgments of trustworthiness can be overridden by experience, even relatively minimal experience. This finding may have important implications for social psychologists interested in stereotype and prejudice. For example, Cunningham et al. (2004) has demonstrated that black faces presented quickly are associated with increased amygdala activation compared to white faces and has suggested that this effect is related to implicit levels of racial bias. However, this effect seems to disappear when the faces are presented for longer lengths of time, presumably when controlled processing can override this automatic evaluation. Our results suggest that repeated positive interactions may also be able to reshape these automatic evaluations. It is important to note, however, that we still observed a significant difference between high trustworthy high reciprocators and low trustworthy high reciprocators, indicating that not all pre-existing judgments were erased by experience. In addition to perceived trustworthiness, we also found that participants believed that the high trustworthy high reciprocators reciprocated at a higher probability compared to the low trustworthy

Please cite this article in press as: Chang, L. J., et al. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology* (2010), doi:10.1016/j.cogpsych.2010.03.001

high reciprocators and the high probability computer controls. Together, these findings indicate that the initial impressions interact with experience when they are congruent to influence both cognition and behavior.

4.3. Modeling trust

To better understand the individual learning processes underlying our behavioral findings, we modeled three possible learning processes. While there are of course many models that could have been tested, we chose to focus on four that have a strong conceptual grounding. First, we tested a pure initialization hypothesis by manipulating the initial starting value based on the normative trust ratings in the GL Initialization model. Next, we tested models that allow the trustworthiness beliefs to influence how feedback is interpreted. The Confirmation Bias model proposes that information consistent with the initial trustworthiness judgment will receive a learning bonus, while inconsistent information (i.e. a high trustworthy face defecting) will be ignored (Biele et al., 2009; Delgado et al., 2005; Doll et al., 2009). The Trust Decay model proposes that initial trustworthiness judgments will influence behavior early in the interactions, but will eventually be overridden by experience, a prediction akin to a dual process model (Frank, O'Reilly, & Curran, 2006; Frank et al., 2007; Poldrack et al., 2001). Finally, the Dynamic Belief model proposes that trustworthiness beliefs consistently serve as a learning bonus in the update function, but are themselves updated based on their partners' behavior after each interaction.

Our cross-validation procedure revealed that a pure initialization account could not explain the data as well as models that allowed the initial beliefs to influence the actual update function. Supporting this finding, we find that our behavioral interaction is not driven by early trials, but rather that it persists until the last trial of the learning sequence (see Fig. 2, Panel B). In addition, our model which allows a learning bonus to only influence the update function during early trials essentially reduced to a constant bonus model, providing additional evidence that this process is not something that is overridden by experience.

Further, we find that a model that allows trustworthiness beliefs to both influence how feedback is interpreted and allow the beliefs to be updated based on this feedback provides the best explanation of the data. Importantly, this model makes two specific predictions that are supported by our data. First, as illustrated in our simulation (Fig. 5), the Dynamic Belief model predicts that by the end of the experiment participants should be investing the most money in trustworthy looking partners that reciprocate frequently and the least amount of money in trustworthy looking partners that reciprocate infrequently. We observed support for this prediction in the last trial of our behavioral data (Fig. 2, Panel B) and moreover observe the exact pattern of behavior for all of the conditions predicted by our model. Second, our Dynamic Belief model predicts that perceptions of trustworthiness will change over time based on actual experiences. This prediction was also supported, as the model-predicted trustworthiness ratings were able to better predict the post-experiment subjective trustworthiness ratings than the initial perceptions of trustworthiness (i.e. the normative trust ratings). Because the model-predicted ratings update based on experience, they were also better able to predict the post-experiment probability ratings than the initial trustworthiness ratings. These findings support and extend previous research, which has found that experience overwhelms description in risky choice (Barron & Erev, 2003; Hertwig, Barron, Weber, & Erev, 2004; Jessup, Bishara, & Busemeyer, 2008). While this notion of dynamically updating beliefs is certainly not new, this study provides, to our knowledge, support for the first computational model of this effect in an iterative social exchange.

5. Conclusion

Our study integrates theories and methods from psychology, economics, and reinforcement learning to gain a greater understanding of how high-level social cues such as trustworthiness are acquired and utilized in a consequential social decision. The findings suggest that trustworthiness judgments may serve as a risk belief (i.e. probability of reciprocation). This belief is based on initial judgments

Please cite this article in press as: Chang, L. J., et al. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology* (2010), doi:10.1016/j.cogpsych.2010.03.001

of perceived trustworthiness and is dynamically updated based on experiences through repeated interactions. This study illustrates the conceptual and methodological advantages of an interdisciplinary approach and provides a novel quantitative framework to conceptualize the notion of trustworthiness as well as an approach to bridge the division between descriptive information and experienced information in the judgment and decision-making literature (Jessup et al., 2008). More broadly, our study provides an important and timely contribution to a growing literature interested in the neural computations underlying social learning (Behrens et al., 2008; Biele et al., 2009; Delgado et al., 2005; Hampton, Bossaerts, & O'Doherty, 2008; King-Casas et al., 2005; Olsson & Phelps, 2007) and trustworthiness (Krueger et al., 2007; Oosterhof & Todorov, 2008; van 't Wout & Sanfey, 2008; Winston, Strange, O'Doherty, & Dolan, 2002) and illustrates the importance of social beliefs in decision-making behavior.

Acknowledgments

The authors thank Niko Warner and Carly Furgersen for their help in the collection of the data and Dr. Mike X. Cohen and three anonymous reviewers for their helpful comments.

References

Adolphs, R., Tranel, D., & Damasio, A. R. (1998). The human amygdala in social judgment. *Nature, 393*(6684), 470–474.
Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.
Apesteguia, J., Huck, S., & Oechssler, J. (2007). Imitation-theory and experimental evidence. *Journal of Economic Theory, 136*(1), 217–235.
Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211*(4489), 1390–1396.
Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 340–412.
Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making, 16*(3), 215–233.
Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., & Fehr, E. (2008). Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron, 58*(4), 639–650.
Behrens, T. E., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. (2008). Associative learning of social value. *Nature, 456*(7219), 245–249.
Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10*, 122–142.
Biele, G., Rieskamp, J., & Gonzalez, R. (2009). Computational models for the combination of advice and individual learning. *Cognitive Science, 33*, 206–242.
Blakemore, S. J., & Decety, J. (2001). From the perception of action to the understanding of intention. *Nature Reviews Neuroscience, 2*(8), 561–567.
Bohnet, I., & Zeckhauser, R. (2004). Trust, risk, and betrayal. *Journal of Economic Behavior and Organization, 55*(4), 467–484.
Camerer, C. (2003). *Behavioral game theory*. New York: Russell Sage Foundation.
Coleman, T. F., & Li, Y. (1996). An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on Optimization, 6*, 418–445.
Cunningham, W. A., Johnson, M. K., Raye, C. L., Chris Gatenby, J., Gore, J. C., & Banaji, M. R. (2004). Separable neural components in the processing of black and white faces. *Psychological Science, 15*(12), 806–813.
Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature, 441*(7095), 876–879.
Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience, 8*(11), 1611–1618.
Doll, B. B., Jacobs, W. J., Sanfey, A. G., & Frank, M. J. (2009). Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Research, 1299*, 74–94.
Frank, R. H., Gilovich, T., & Regan, D. T. (1993). The evolution of one-shot cooperation: An experiment. *Ethology and Sociobiology, 4*(4), 247–256.
Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences of the United States of America, 104*(41), 16311–16316.
Frank, M. J., O'Reilly, R. C., & Curran, T. (2006). When memory fails, intuition reigns: Midazolam enhances implicit inference in humans. *Psychological Science, 17*(8), 700–707.
Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences of the United States of America, 105*(18), 6741–6746.
Hampton, A. N., & O'Doherty, J. P. (2007). Decoding the neural substrates of reward-related decision making with functional MRI. *Proceedings of the National Academy of Sciences of the United States of America, 104*(4), 1377–1382.
Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior, 4*, 267–272.
Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15*(8), 534–539.

Please cite this article in press as: Chang, L. J., et al. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology* (2010), doi:10.1016/j.cogpsych.2010.03.001

Jessup, R. K., Bishara, A. J., & Busemeyer, J. R. (2008). Feedback produces divergence from prospect theory in descriptive choice. *Psychological Science, 19*(10), 1015–1022.
Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–291.
King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. *Science, 308*(5718), 78–83.
Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature, 435*(7042), 673–676.
Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R., Strenziok, M., et al (2007). Neural correlates of trust. *Proceedings of the National Academy of Sciences of the United States of America, 104*(50), 20084–20089.
Morris, J. S., Friston, K. J., Buchel, C., Frith, C. D., Young, A. W., Calder, A. J., et al (1998). A neuromodulatory role for the human amygdala in processing emotional facial expressions. *Brain, 121*(Pt 1), 47–57.
Olsson, A., & Phelps, E. A. (2007). Social learning of fear. *Nature Neuroscience, 10*(9), 1095–1102.
Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America, 105*(32), 11087–11092.
Pizzagalli, D. A., Lehmann, D., Hendrick, A. M., Regard, M., Pascual-Marqui, R. D., & Davidson, R. J. (2002). Affective judgments of faces modulate early activity (approximately 160 ms) within the fusiform gyri. *Neuroimage, 16*(3 Pt 1), 663–677.
Poldrack, R. A., Clark, J., Pare-Blagoev, E. J., Shohamy, D., Moyano, J. C., Myers, C., et al (2001). Interactive memory systems in the human brain. *Nature, 414*, 546–550.
Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology, 49*, 95–112.
Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II* (pp. 64–99). Appleton-Century-Crofts.
Scharlemann, J. P. W., Eckel, C. C., Kacelnik, A., & Wilson, R. K. (2001). The value of a smile: Game theory with a human face. *Journal of Economic Psychology, 22*, 617–640.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.
Simpson, J. A. (2007). Psychological foundations of trust. *Current Directions in Psychological Science, 16*(5), 264–268.
Singer, T., Kiebel, S. J., Winston, J. S., Dolan, R. J., & Frith, C. D. (2004). Brain responses to the acquired moral status of faces. *Neuron, 41*(4), 653–662.
Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*(2), 245–251.
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
van 't Wout, M., & Sanfey, A. G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition, 108*(3), 796–803.
Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*(7), 592–598.
Winston, J. S., Strange, B. A., O'Doherty, J., & Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience, 5*(3), 277–283.
Yechiam, E., Busemeyer, J. R., Stout, J. C., & Bechara, A. (2005). Using cognitive models to map relations between neuropsychological disorders and human decision-making deficits. *Psychological Science, 16*(12), 973–978.
Zak, P. J., & Knack, S. (2001). Trust and growth. *The Economic Journal, 111*(470), 295–321.

Please cite this article in press as: Chang, L. J., et al. Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology* (2010), doi:10.1016/j.cogpsych.2010.03.001