So good it has to be true: Wishful thinking in theory of mind

Daniel Hawthorne-Madell (djthorne@stanford.edu), **Noah D. Goodman** (ngoodman@stanford.edu)
Department of Psychology, Stanford University Stanford, CA USA 94350

Abstract

In standard decision theory, rational agents are objective, keeping their beliefs independent from their desires (Berger, 1985). Such agents are the basis for current computational models of Theory of Mind (ToM), but this fundamental assumption of the theory remains untested. Do people think that others' beliefs are objective, or do they think that others' desires color their beliefs? We describe a Bayesian framework for exploring this relationship and its implications. Motivated by this analysis, we conducted two experiments testing the *a priori* independence of beliefs and desires in people's ToM and find that, contrary to fully-normative accounts, people think that others engage in *wishful thinking*. In the first experiment, we found that people think others believe both that desirable events are more likely to happen, and that undesirable ones are less likely to happen. In the second experiment, we found that social learning leverages this intuitive understanding of wishful thinking: participants learned more from the beliefs of an informant whose desires were contrary to his beliefs. People's ToM therefore appears to be more nuanced than the current rational accounts, but consistent with a model in which desire directly affects the subjective probability of an event.

Keywords: Wishful Thinking; Computational Social Cognition; Theory of Mind; Desirability Bias

Introduction

"I think Romney will take Ohio" Karl Rove intoned into the camera while Fox News' election experts were calling Ohio, and the 2012 election, for Barack Obama. The strength of Mr. Rove's desire to see a Romney victory was palpable and it seemed to overpower the evidence from the exit polls to form his belief. However, this explanation presupposes a direct link between his desires and beliefs, a link that is currently absent in normative behavioral models and current Theory of Mind (ToM) models.

Does a causal link between desires and beliefs exist?[1] The evidence is mixed. There are a number of compelling studies that find "wishful thinking," or a "desirability bias" in both carefully controlled laboratory studies (Mayraz, 2011) and real world settings, such as the behavior of sport fans (Babad, 1987; Babad & Katz, 1991), expert investors (Olsen, 1997), and voters (Redlawsk, 2002). However, other researchers have failed to observe the effect, e.g., Bar-Hillel et al.'s *The elusive wishful thinking effect* (1995), have provided alternative accounts of previous experiments (Hahn & Harris, 2014), and have argued that there is insufficient evidence for a systematic wishful thinking bias (Krizan & Windschitl, 2007; Hahn & Harris, 2014).

Whether or not there actually *is* a direct effect of desires on beliefs, people might *think* that there is and use this fact when reasoning about other people. That is to say, people's ToM might include this causal influence (as seen in Fig. 1b).

[1] While the causal link between desires and beliefs may, in fact, be bi-directional, we will focus on the evidence for the *a priori* effect of desires on beliefs.

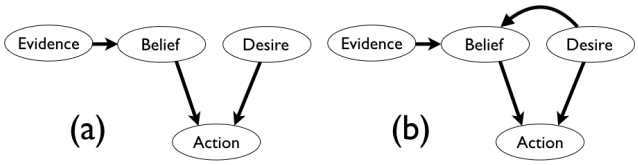

Figure 1: Causal models of (a) ToM based upon classic belief-desire psychology and (b) ToM that includes a direct "wishful thinking" link between desires and beliefs.

The direct influence of desires on beliefs is a departure from classic belief-desire "folk" psychology in which beliefs and desires are independent and jointly cause action (Fig. 1a). Previous models of ToM formalize belief-desire (B-D) psychology into generative models of action and belief formation. They show that inferring others' beliefs (Baker, Saxe, & Tenenbaum, 2011), preferences (Jern, Lucas, & Kemp, 2011), and desires (Baker, Saxe, & Tenenbaum, 2009) can be understood as Bayesian reasoning over these generative models. A fundamental assumption of these models, and the B-D theory underpinning them, is that beliefs are formed on the basis of evidence, and *a priori* independent of desire. We will call models that make this assumption *rational theories of mind* (rToM). We can contrast this rationally motivated theory with one that incorporates the rose colored lenses of a desire-belief connection, an *optimistic ToM* (oToM).

In this paper we attempt to experimentally distinguish rToM from oToM. This is difficult in typical social reasoning tasks because actions are observed, which renders desire and belief conditionally dependent, even in rToM. To isolate the *a priori* relation we therefore perform two experiments in which participants judge the likely beliefs of an agent who has not taken an action, in situations where the agent's likely desires and evidence vary. We begin by formalizing rToM and oToM as probabilistic models in order to make their predictions more explicit.

Models of belief formation and social learning

The standard rational theory of mind pictured in Figure 1a postulates a theory of belief formation that can be understood as Bayesian updating of beliefs about unobserved states given some observed evidence (and a causal model of the world). For concreteness, we describe the models in this section in terms of the scenario employed in Experiment 2. A fan forms belief about their team's skill, b, by updating after seeing evidence, e, in the form of the team's record in their past matches, $\mathcal{E}$. By Bayes rule:

$$p(b|e,\mathcal{E}) \propto p(e|\mathcal{E},b)p(b), \tag{1}$$

where $p(b)$ is the prior probability of the team having skill b and $p(e|b,\mathcal{E})$ is the likelihood that the team won e of the $\mathcal{E}$ matches given skill b. We assume skill is simply the probability that the team will win a match (and therefore we can talk of a team's skill and chance of winning the next match interchangeably). Given this assumption, each match is independent and can be predicted by flipping a coin weighted by the team's skill: $e \sim \text{Binomial}(b,\mathcal{E})$.

In the rational theory of mind described in Eq. 1, the fan's belief in the team's skill (b) is *a priori* independent of their desire.[2] The "optimistic" theory of mind pictured in Figure 1b breaks this independence assumption, with beliefs directly depending upon desires. We formalize belief update in this model by:

$$p(b|e,\mathcal{E},d) \propto p(e|\mathcal{E},b,d)p(b|d)p(d) \quad (2)$$
$$\propto p(e|\mathcal{E},b)p(b|d), \quad (3)$$

where the second line follows by assuming that evidence doesn't change the agent's desires, $e \perp d|b$, and assuming a uniform $p(d)$ for simplicity. To capture *wishful* thinking, we assume that the direction of $p(b|d)$ is for positive utility desires to lead to higher prior probability for the corresponding events; for example, passionate fans are more likely to think their team will win.

The difference between belief update in rToM and oToM can therefore be understood as a dependence of prior belief on desire. A reasoner using a rToM (Eq. 1) would infer the same belief for a person who saw evidence e regardless of whether the person desired or dreaded the event (see Fig. 3d). However, a reasoner using an oToM (Eq. 3) would infer a person believes desired events to be more probable than dreaded ones (see Fig. 3c).

Reasoners using these different theories of mind not only differ in the inferences they draw about others' beliefs but also in how they can use knowledge of others' beliefs to make inferences about the world, i.e., social learning. This can be seen by considering a fan who passes a signal s to the reasoner that indicates whether they think their team will win: $p(s|b) = \delta_{b \geq 0.5}$. In Experiment 2, reasoners don't know what the fan believes about a team's strength, and instead have to infer it from their own prior beliefs about the team's skill $p_r(b)$, the matches the fan saw $\mathcal{E}$, and the fan's desire d. To infer the true team strength (which we subscript as b_r for clarity) the reasoner has to consider what evidence e the fan actually saw:

$$p(s|b_r,\mathcal{E},d) = \sum_e p(s|b_r,\mathcal{E},e,d)p(e|b_r,\mathcal{E},d) \quad (4)$$
$$= \sum_e p(s|\mathcal{E},e,d)p(e|b_r,\mathcal{E}), \quad (5)$$

where the second line follows from the fact that $s \perp b_r|e$ (the fan's signal depends on the true strength only via the evidence) and $e \perp d$ (desire doesn't influence the actual evidence). Eq. 5 represents the probability that the fan would send signal s given the true team skill. We take $p(e|b_r,\mathcal{E})$ to be binomial as above. The remaining term can be expanded to make the fan's belief explicit:

$$p(s|e,\mathcal{E},d) = \int_b p(s|b,d,e,\mathcal{E})p(b|e,\mathcal{E},d)\ \mathrm{d}b \quad (6)$$
$$= \int_b p(s|b)p(b|e,\mathcal{E},d)\ \mathrm{d}b \quad (7)$$

which depends on the belief formation ($p(b|e,\mathcal{E},d)$) and signaling ($p(s|b)$) models specified above. Finally, using Eq. 5, we can describe the reasoner who learns about a team's skill from social information—a fan's desire, amount of evidence, and signal. Using Bayes rule:

$$p(b_r|s,\mathcal{E},d) \propto p(s|b_r,\mathcal{E},d)p(b_r) \quad (8)$$

Eq. 8 represents a social learner who assumes an oToM; for a social learner who assumes a rToM the dependence on d disappears.

The qualitative behavior of the rational and optimistic social learning models can be seen in Figure 4b and c. A reasoner using a rToM considers only the amount of evidence a fan seen, $\mathcal{E}$, i.e., the fan's knowledgeability (for a given set of direct evidence e_r). For a reasoner using an oToM, the desirability of the outcome influences their estimate. Given the *a priori* bias to believe in desirable events, when a person desires an outcome and yet believes that it will not occur, a reasoner can infer that they saw *strong evidence* that the outcome will not occur. The oToM reasoner therefore learns more from this person than they would from someone who had identical beliefs that were consistent with their desires.

We have outlined the patterns of reasoning expected if reasoners use an oToM when thinking about others. Whether people actually conform with these predictions, assuming that beliefs are *a priori* dependent on desires, is an open empirical question. We therefore conducted two experiments testing the qualitative predictions of oToM. In Experiment 1, we test whether reasoner's inferences about others' beliefs reflect an *a priori* dependence of beliefs on desires. A consequence of such a dependence would be that learners are sensitive to others' desires when learning from them, which we explore in Experiment 2.

Given the strongly divergent predictions of the optimistic and rational ToM models, only a qualitative comparison is needed to show the presence and consequences of wishful thinking in ToM. To generate the qualitative predictions in Figure 3c and 4b, we used equation Eq. 3 and 8 where we defined wishful thinking as a prior biased in the direction of the desire; we assumed that a belief b (given a desire d) was drawn from a Beta distribution whose mean was biased towards the desired outcome with the magnitude of this bias representing the degree of wishful thinking. We fit the mean and the variance of this Beta distribution to the data in each experiment.

Experiment 1: Wishful thinking in ToM

To test for the presence of wishful thinking in people's mental models of others we introduce Josh, a person playing a

[2] Although beliefs and desires are conditionally dependent given an action in rToM.

game with a transparent causal structure. The causal structure of the game is conveyed via the physical intuitions of the Galton board pictured in Fig. 2c. The outcome of the game is binary (there are two bins) with different values associated with each outcome (money won or lost). We call the value of an outcome (i.e., the amount that Josh stands to win or lose) the utility of that outcome, $U(outcome)$. Participants are asked what they think about Josh's belief in the likelihood of the outcome $p_J(outcome)$. By manipulating outcome values we are able to test for wishful thinking. If people incorporate wishful thinking into their ToM, we should find that increasing an outcome's utility ($U(outcome)$) results in higher estimates of Josh's belief in the outcome's occurrence ($p_J(outcome)$).

We first measured $p_J(outcome|evidence)$ without manipulating the desirability of the outcome in the "baseline" block of trials. Then in the "utility" block of trials we assigned values to outcomes, manipulating Josh's $U(outcome)$.[3] In the *utility* block of trials we used a *Price Is Right*-style spinning wheel (Fig. 2a and b) to show Josh (and the participant) what he stood to win or lose based on the outcome of the marble drop. By comparing these two blocks of trials we test for the presence of wishful thinking in people's ToM.

Methods

Participants We recruited 110 participants via Amazon Mechanical Turk and paid them $.75. Participants were split into into two conditions: the *Dual outcome* (25 male and 20 female, μ_{age} = 28, σ_{age} = 9.1) and *Many outcome* (31 male and 34 female, μ_{age} = 27, σ_{age} = 8.9).[4] Ten participants were excluded from the analyses for responding incorrectly to attention checks.

Design and Procedure

Participants first were introduced to Josh[5] who was playing a marble-drop game with a Galton board (as seen in Figure 2c). Josh was personified as a stick figure and appeared on every screen. To provide participants with an example of the causal structure (i.e., physics) of the game, they were first shown a marble dropping from the center of the board, twice. One marble landed in the orange bin (Figure 2c left bin) and the other landed in the right (Figure 2c right bin). After observing the physical properties of the board (i.e., the two marble drops) participants began the *baseline* block of trials. In the four baseline trials, the marble's drop position varied and participants were asked "What do you think Josh thinks is the chance that the marble lands in the bin with the purple/orange box?" Participants' responses were recorded on a continuous

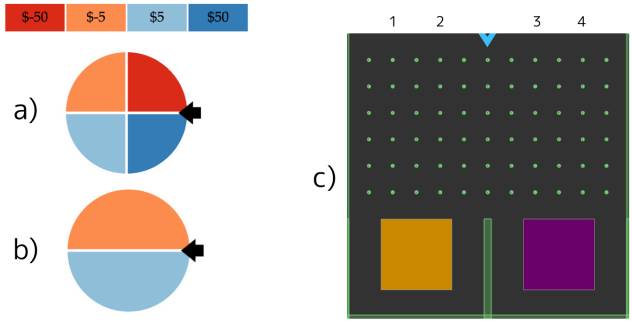

Figure 2: Stimuli used in Experiment 1. (a) the wheel used to determine the payout for the next outcome in the *Many outcome* condition and (b) in the *Dual outcome* condition. (c) Galton board used to decide the outcome in Experiment 1. The blue arrow at the top indicates where the marble will be dropped. The numbers indicate the four drop positions used in the experiment.

slider with endpoints labeled "Certainly Will" and "Certainly Won't." Color placement was randomized on each trial, and the color of the box in question varied between participants. The marble drop position was indicated with a blue arrow at the top of the Galton board and there were four drop positions used ($marble_x$; top of Figure 2c) which varied in how likely they were to deliver the marble into the bin in question.

After the baseline trials, participants were introduced to the *utility* trials, which included a spinning wheel labeled with outcome values that determined "how much Josh can win or lose." In the Dual outcomes condition, Josh could win or lose $5 (as seen in 2a), and in the Many outcomes condition, Josh could win or lose $5 or $50 (as seen in 2b). At the beginning of each trial the wheel was spun and the selected payout was displayed, e.g., "Josh has a chance of winning $5," along with the Galton board. The bins were labeled with a $ and ∅ symbol.[6] If the marble landed in the $ bin then Josh won/lost the money. The location of the $ bin was randomized on each trial. After seeing the Galton board with $marble_x$ indicated with a blue arrow, participants were asked two questions sequentially. First they were asked "What do you think Josh believes is the chance that the marble will land on the {$/-$} and he'll {win/lose} {$5/ $50}," with the response recorded on the same slider as the baseline trials with endpoints labeled "Certainly Will" and "Certainly Won't." They were then asked "How much does Josh care about the outcome?" with the response on a slider with endpoints labeled from "Not at All" to "To a Great Extent."

In the Dual outcome condition, participants saw every combination of the two outcomes ($5, -$5) and the four drop positions (20%, 40%, 60%, and 80%) for a total of 8 utility trials. In the Many outcome condition participants saw 8 random samples from the 16 possible combinations of the four payouts and the four drop points. Each participant also saw 5 catch trials asking either where the marble had landed after

[3] Crucially, Josh's $U(outcome)$ should not be chosen by him, e.g., "I bet $5 that it lands in the right bin," as such an action would render $U(outcome)$ and $p(outcome)$ conditionally dependent and both rToM and oToM would predict influence of desire on belief judgments. To test pure wishful thinking, Josh's $U(outcome)$ has to be assigned to him by a process independent of $p(outcome)$.

[4] These two conditions were presented as two separate HITs on Amazon Mechanical Turk, two weeks apart, with no participant allowed to participate in both conditions.

[5] A random male name was generated for each participant.

[6] $ when the payout was positive and -$ when it was negative.

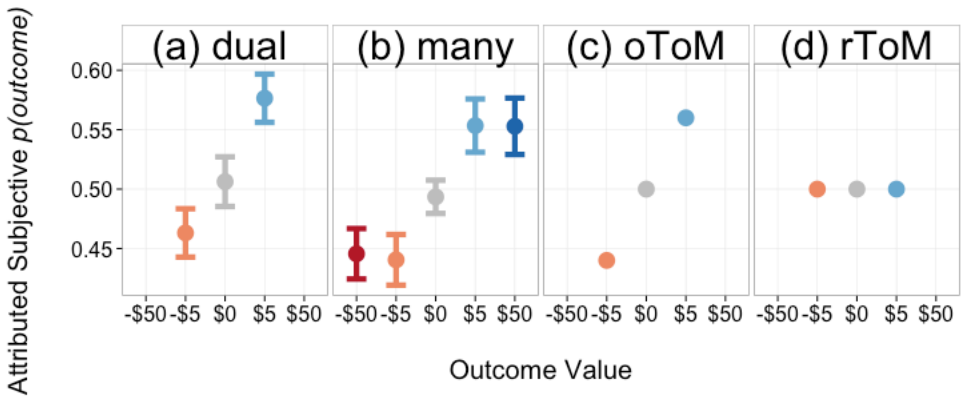

Figure 3: For each outcome value, the mean subjective $p(outcome)$ attributed to the agent is shown (with standard error bars) for (a) the Dual outcome condition and (b) the Many outcome condition. These data are compared with the qualitative predictions of the (c) optimistic and (d) rational ToM models.

a Galton board demonstration, or comprehension questions about the game and their current task.

Results and Discussion

Comparison of Dual and Many Condition As evident in Figure 3b, participants showed no sensitivity to the magnitude of the value of the outcome. We therefore combined $5 and $50 into one positive-value categorical variable and -$5 and -$50 into a negative-value categorical variable and tested whether this qualitatively coded Many condition data differed from the Dual condition data. For each level of value—negative, baseline, and positive—we compared the average reported $p(outcome)$ between the Dual and Many conditions by permutation test. The tests indicated that the results are not distinguishable ($p > .05$ for each level of value). In the subsequent analyses we therefore combine the results from the two conditions.

Wishful thinking in ToM In a rational theory of mind, beliefs and desires are *a priori* independent. Manipulating Josh's desires therefore shouldn't have an effect on his beliefs, and we would predict that the utility trials look like the baseline trials. However, as seen in Figure 3a and b, the utility trials varied systematically from the baseline trials and the predictions of a rToM. To quantify this deviation we fit a linear mixed effects model to participants' $p(outcome)$ responses. The model used $marble_x$ and the categorically coded value of the outcome (negative, baseline, and positive) as fixed effects and included the random effect of $marble_x$, outcome value, and intercept for each participant. The resulting model indicated that if an outcome was associated with a utility for Josh, participants thought that it would impact his beliefs about the probability of that outcome. Participants' thought that Josh would believe that an outcome that lost him money was less likely than the corresponding baseline trial ($\beta = -.047, t(98) = -3.7, p < .05$).[7] They also thought that an outcome that would net him money was more likely than the corresponding baseline trial ($\beta = .064, t(97) = 5.5, p < .001$).[8] Finally, $marble_x$, the direct evidence, had a significant influence ($\beta = .79, t(99) = 18, p < .05$). There was no evidence that the effect of the outcome value was affected by $marble_x$ (the interactive model provided a worse fit ($\chi^2(1) = .82, p < .05$).

The results from Experiment 1 are consistent with the qualitative predictions of the oToM model (Eq. 2) indicating that people's ToM includes a direct "wishful thinking" link, unlike rToM models where beliefs and desires are *a priori* independent. To test the robustness of this finding, in Experiment 2 we expand our sights to social learning situations where oToM (but, crucially, not rToM) predicts that desires affect a social source's influence.

Experiment 2: Learning from others with an oToM

Do people consider a social source's desires when learning from him? It would be important to do so if they think that his desires have a direct influence on his beliefs. Consider a learner using an oToM to reason about her uncle, a Cubs fan, who proudly proclaims that this is the year the Cubs win the pennant[9]. Though her uncle knows a lot about baseball, the oToM learner is largely unmoved from her (understandably) skeptical stance. However, if her aunt, a lifelong Yankees fan, agrees that the Cubs do look better than the Yankees this year, then an oToM learner considers this a much stronger teaching signal. A learner reasoning with a rToM wouldn't distinguish between these two social sources[10] as seen in Figure 4c.

[7] Denominator degrees of freedom used to calculate p-values were approximated using the Satterthwaite method

[8] There was no evidence of loss aversion in the relative magnitude of the wishful thinking effect for positive and negative utilities. In fact, the magnitude of the wishful thinking effect was slightly stronger for positive utilities.

[9] It never is.

[10] Assuming that his aunt and uncle are equally knowledgeable and their statements have no causal influence on the game—if the uncle is a referee, his desires may matter through more objective routes.

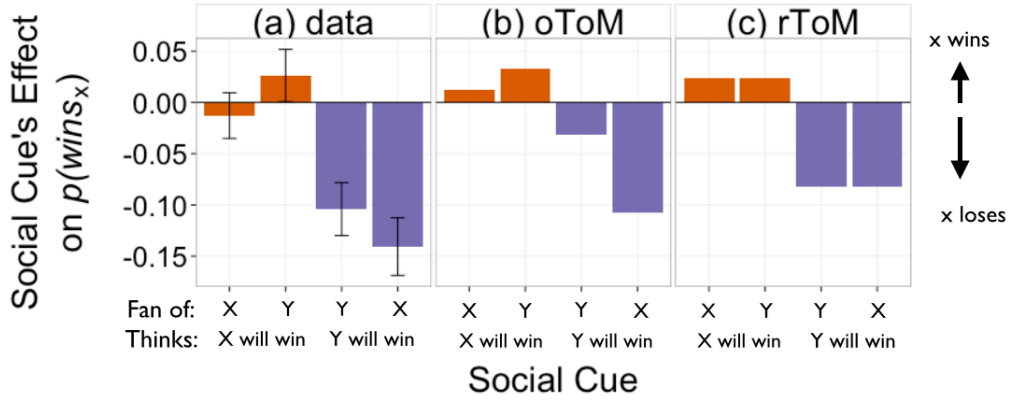

Figure 4: Effect of a social sources' desire on how others learn from them for (a) Experiment 2 data with standard error bars, which we compare to the qualitative predictions of (b) an optimistic ToM, (c) a rational ToM. Bars represent the difference between the baseline mean $p(win_x)$ response and the mean $p(win_x)$ response for the four social trials (where $Team_x$ is observed to win 3 of the 5 last matches). The color of the bars corresponds to which team the fan thought would win. The difference between each pair of colored bars therefore represents the effect of the fan's desire.

We investigated which ToM best describes learning from social sources using a scenario where participants reasoned about the upcoming match of a fictional soccer tournament. Participants were introduced to a fan from one of the schools who said which team they thought would win. Combining the two possible desires ($d_f \in \{win_x, win_y\}$) with the two possible beliefs a fan could have ($b_f \in \{win_x, win_y\}$) yields four possible social cues. In addition to the social cues, participants were given direct evidence of $p(win_x)$ in the form of previous match outcomes, to make questions about the next outcome more natural. The effect of the social information was isolated by comparing the social trials to corresponding baseline trials that only provided direct evidence.

Participants

40 participants (22 female, μ_{age} = 31.4, σ_{age} = 9.4) were recruited via Amazon Mechanical Turk and paid $.60. Five participants were excluded from the analyses for failing to respond to attention checks.

Design Procedure

Participants were told about a (fictional) annual British collegiate soccer tournament where teams played one another year after year in the initial round robin phase, creating rivalries. Participants were given some direct evidence of each rivalry in the form of a summary table of the results of their last five matches. In the initial *baseline* trials, participants just received this non-social, direct evidence about each rivalry and were asked, "What do you think is the chance that $Team_x$ wins the match this year?" They were presented four baseline trials such that they saw $Team_x$ win $1/5$, $2/5$, $3/5$, and $4/5$. The names for $Team_x$ and $Team_y$ were randomly selected from a database of British counties and assigned a random color which was shown as a border around the team's name in the results table.The $2/5$, $3/5$ trials were used as reference baselines for the impact of the corresponding social trials.

Following the baseline trials, participants were introduced to a "student of one of the colleges who is a big fan of his school's team ...(who will) say who he thinks will win this year's game." Each social trial consisted of a cartoon of the student wearing his team's (randomly assigned) color. He introduces himself, "I've seen the last 10 matches of this rivalry because I'm a big fan of {Team}" and then either professed to be bullish("and I think they will win this year") or bearish ("but I think {Other Team} will win this year") on his team. Each trial also included a results table where $Team_x$ won 3 out of 5 times.[11] Participants were then asked the same question as in the baseline trials "What do you think is the chance that $Team_x$ wins the match this year." Participants saw four social trials, one for each combination of who the fan believed will win and who he wanted to win.

Each participant saw 4 catch trials asking either what the results of the last match up where, or comprehension questions about the game and their current task.

Results and Discussion

As seen in Figure 4a, participants' estimates of $p(win_x)$ reflect the fan's beliefs, b_s, as predicted by both the rToM and the oToM model. However, contrary to the predictions of rToM, the effect of a fan's beliefs (b_s) appears to depend upon his desires (d_s). To quantify this effect we fit a linear mixed effects model to participants' $p(win_x)$ responses using b_s, d_s, and their interaction as fixed effects in addition to the random effect of b_s and intercept for each participant (there were insufficient within-participant data to estimate additional random effect parameters). The interactive mixed model provides a significantly better fit compared to

[11] this was counterbalanced such that half of the time the team that was the subject of the question won $2/5$, but we will talk about the trials in the 'canonical form' where the participant is asked about $Team_x$ who won $3/5$.

a model that just includes the additive effects of b_s and d_s ($\chi^2(1) = 4.9, p < .05$). Both the main effect of b_s and its interaction with d_s was significant in the interactive mixed model ($t(64) = -5.7, p < .05$ and $t(62) = 2.3, p < .05$, respectively).

As seen in Figure 4a and b, the results are consistent with the qualitative predictions of the model of learners that uses an oToM, where the fan's desires have a direct influence on their beliefs (Eq. 8). The influence of equally knowledgeable fans who expressed the same beliefs depended on what the fans wanted to happen. Looking at the red bars in Figure 4a, we see that fans that believed the more likely team would win changed participants' judgments more when this belief ran against their desire. This is to say that people do learn more from agents who believe things that are contrary to their desires as predicted by the model of oToM learners.

Discussion and Conclusion

Current computational models of theory of mind are built upon the assumption that beliefs are *a priori* independent of desires. Whether social reasoners use such a rational ToM (rToM) is an empirical question. In two experiments we tested the independence of beliefs and desires in ToM and found systematic evidence that people think that others are wishful thinkers whose beliefs are colored by their desires. In Experiment 1 we found that people believe that others inflate the probability of desirable outcomes and underestimate the probability of undesirable ones, as an optimistic ToM (oToM) with a direct link between desires and beliefs would predict. Our model results predicted that if social learners used an oToM to reason about others, we should expect their learning to be affected by the desires of these social sources. Indeed, in Experiment 2 we found that learners were more influenced by sources whose beliefs ran against their desires. Taken together these experiments suggest that people have a nuanced ToM, with systematic deviations from the rational B-D psychology underpinning rToM. However, further investigations are required to show that people spontaneously employ an oToM when desirability is manipulated between-subjects, and therefore less salient.

The presence of wishful thinking in ToM has no necessary relation to its existence in human "online" reasoning under uncertainty. Indeed, the considerable heterogeneity of the wishful thinking effect discussed in the literature leaves open the possibility that people could think that others' desires are coloring their beliefs when, in fact, they are not. If this were the case, it could help explain why first-person wishful thinking is reliably found in some paradigms and not others. The paradigms in which wishful thinking is reliably found involve participants *reasoning about themselves and others* (for a review see Shepperd, Klein, Waters, & Weinstein, 2013), whereas *asocial paradigms* involving direct estimation of probabilities do not find the effect (e.g., Bar-hillel & Budescu, 1995). Experiment 1 provides the opportunity for an additional test of this explanation, comparing the current results to the experiment framed as a task in which there is no other agent and participants themselves stand to win.

The experiments presented here suggest that people think that others are wishful thinkers; this has broad consequences for social reasoning ranging from our inferences about pundit-posturing to self-regulation. Our findings highlight the importance of further research into the true structure of ToM. Do people think that others exhibit loss aversion or overweight low probabilities? Is the connection between beliefs and desires bi-directional? Rigorous examination of questions like these may buttress new, empirically motivated computational models of ToM that capture the nuance of human social cognition—an idea so good it has to be true.

Acknowledgments

This work was supported by an NSF Graduate Research Fellowship, ONR grant N000141310341, James S. McDonnell Foundation Scholar Award to NDG. We thank Long Ouyang and Gregory Scontras for helpful feedback.

References

Babad, E. (1987). Wishful thinking and objectivity among sports fans. *Social Behaviour*, *2*(23).

Babad, E., & Katz, Y. (1991). Wishful Thinking-Against All Odds. *Wishful thinking-Against all odds. Journal of Applied Social Psychology*, *21*, 1921–1938.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society*.

Bar-hillel, M., & Budescu, D. (1995, September). The elusive wishful thinking effect. *Thinking and Reasoning*, *1*.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (Vol. 83). Springer.

Hahn, U., & Harris, A. J. L. (2014). *What Does It Mean to be Biased: Motivated Reasoning and Rationality*.

Jern, A., Lucas, C. G., & Kemp, C. (2011). Evaluating the inverse decision-making approach to preference learning. *NIPS*, 2276–2284.

Krizan, Z., & Windschitl, P. D. (2007). The influence of outcome desirability on optimism. *Psychological bulletin*, *133*(1), 95–121.

Mayraz, G. (2011). Wishful Thinking. *CEP Discussion Paper*, *1092*.

Olsen, R. A. (1997). Desirability bias among professional investment managers: some evidence from experts. *Journal of Behavioral Decision Making*, *10*(1), 65–72.

Redlawsk, D. P. (2002, November). Hot Cognition or Cool Consideration? Testing the Effects of Motivated Reasoning on Political Decision Making. *The Journal of Politics*, *64*(04), 1021–1044.

Shepperd, J. a., Klein, W. M. P., Waters, E. a., & Weinstein, N. D. (2013). Taking Stock of Unrealistic Optimism. *Perspectives on Psychological Science*, *8*(4), 395–411.