Modeling Self-Deception within a Decision-Theoretic Framework

Jonathan Y. Ito, David V. Pynadath, and Stacy C. Marsella

Information Sciences Institute
University of Southern California
4676 Admiralty Way, Marina del Rey CA 90292 USA

Abstract. Computational modeling of human belief maintenance and decision-making processes has become increasingly important for a wide range of applications. In this paper, we present a framework for modeling the human capacity for self-deception from a decision-theoretic perspective in which we describe processes for determining a desired belief state, the biasing of internal beliefs towards the desired belief state, and the actual decision-making process based upon the integrated biases. Furthermore, we show that in some situations self-deception can be beneficial.

1 Introduction

A mother has been shown seemingly incontrovertible evidence of her son's guilt. Although the information is provided by reliable sources, the mother continues to proclaim her son's innocence. This illustrates an important characteristic of human belief maintenance: that our beliefs are not formed merely by the evidence at hand. Rather, desires and intentions interfere with the processes that access, form and maintain beliefs and thereby bias our reasoning.

Research on human behavior has identified a range of rational as well as seemingly irrational tendencies in how people manage their beliefs [10]. Research in human emotion has detailed a range of coping strategies such as denial and wishful thinking whereby people will be biased to reject stressful beliefs and hold on to comforting ones [11]. Research on cognitive dissonance [8] has demonstrated that people often seek to achieve consistency between their beliefs and behavior. Specifically, cognitive dissonance research has especially focused on how we alter beliefs in order to resolve inconsistencies between a desired positive self-image and our behavior [1], much like Aesop's fable of the fox and the grapes in which after repeatedly failing to reach a bunch of grapes the fox gives up and concludes that the grapes did not look so delicious after all. Similarly, research has also shown a tendency for what is called motivated inference, the tendency to draw inferences and therefore beliefs, based on consistency with one's motivations as opposed to just the facts. Research on how people influence each other has also identified a range of influence tactics that are not simply based on providing factual evidence. However, these are not unconstrained; people do not, cannot, simply believe whatever they choose.

Computational modeling of these human belief maintenance mechanisms has become important for a wide range of applications. Work on virtual humans and Embodied Conversational Agents increasingly has relied on research in modeling human emotions and coping strategies to create more life-like agents [9]. Work in agent-based modeling of social interaction has investigated how persuasion and influence tactics [4] can be computationally modeled [14] for a variety of applications such as health interventions designed to alter user behavior [3].

In this work, we approach the issue of human belief maintenance from the perspective of decision-theoretic reasoning of agents in a multi-agent setting. Specifically, we argue that a range of self-deceptive phenomena can be cast into a singular framework based upon Subjective Expected Utility (SEU) Theory. To cast the seemingly irrational process of wishful thinking and self-deception into a decision-theoretic framework may in itself seem irrational. However, we argue that seemingly irrational behavior such as wishful thinking, motivated inference, and self-deception can be grounded and integrated within an agent's expected utility calculations in a principled fashion.

2 Self Deception Framework

Psychological literature on self-deception commonly refers to the *act* of self-deception as the internal biasing processes involved in adopting a desired belief in the face of possibly contradictory evidence [5, 16]. Therefore much of this literature focuses primarily on these biasing processes and oftentimes the definition of the desired belief state itself remains very abstract. However, by employing Utility Theory in general and SEU-Theory in particular, we are provided a means by which to not only bias beliefs towards a desired belief state and thus influence the subsequent decision-making process but also to designate the desired belief state itself given the decision-maker's own preferences.

SEU-Theory provides the basis for our formulation of self-deception. With it, we not only are able to define the final decision-making process, but also derive the desired belief state of a self-deceptive individual. SEU-Theory as defined by Savage [19] mathematically quantifies an individual's subjective preferences by assigning a numerical utility value to acts performed in a given state. More concretely, SEU is defined as the following equation in which a is some available action, S the set of possible states, $p(s)$ is the probability of state s occuring, and $\mu(a, s)$ is the utility of performing action a in state s:

$$SEU(a) = \sum_{s \in S} p(s)\, \mu(a, s) \tag{1}$$

2.1 Desired Belief Formulation

The existence and specification of the desired belief state is essential to the subsequent biasing procedure of our self-deceptive process. And since SEU-Theory provides a representation of desires based on the utilities of the decision-maker

it serves as an appropriate platform for the operationalization of the desired belief specification process. Just as SEU-Theory maximizes the expected utility over *actions* we will define a similar process that maximizes expected utility over *beliefs* which we define as the Subjective Expected Belief Utility (SEBU).

The SEBU of a particular belief is the SEU a decision-maker can expect assuming that the belief in question is accurate. Formally, the SEBU is evaluated as follows in which $p_b(s)$ is the probability according to belief b of state s occurring and a_b is the action which would be chosen according to SEU-Theory under belief b:

$$SEBU(b) = \sum_{s \in S} p_b(s)\, \mu(a_b, s) \tag{2}$$

Alternatively, we can explicitly include the selection of action a_b in our equation and redefine SEBU as follows:

$$SEBU(b) = \max_{a \in A} \left(\sum_{s \in S} p_b(s)\, \mu(a, s) \right) \tag{3}$$

The selection process of a desired belief is akin to SEU-Theory's expected utility maximization process and is defined as:

$$P' = \operatorname*{argmax}_{b \in B} SEBU(b) \tag{4}$$

We now illustrate the desired belief formulation process with a simple example:

Example 1. Let us revisit the example of a mother proclaiming her son's innocence despite iron-clad evidence to the contrary. We can represent the mother's dilemma as a simple decision problem consisting of 2 states as shown in Table 1. Furthermore we make the assumption that the best possible outcome, with respect to the mother's preferences, is a steadfast belief of her son's innocence coinciding with actual innocence. We also assume that the worst possible outcome is a belief in her son's guilt when in actuality he is innocent. After making these assumptions, only 2 other possible preference orderings remain: $a \succ b \succeq c \succ d$ or $a \succ c \succeq b \succ d$. The former preference ordering is one in which the mother will always choose to proclaim her son's innocence regardless of the evidence presented. And since this behavior is coincidental with the mother's desired belief that her son is innocent, let us instead consider the preference ordering of $a \succ c \succeq b \succ d$. To illustrate the process of desired belief formulation we assign numerical utilities to the various outcomes in accordance with our preference ordering as seen in Table 2. Furthermore, consider the three candidate belief distributions: b_0 in which there is an equally likely probablity that the son is innocent or guilty, b_1 where the son is certainly innocent, and b_2 where the son is certainly guilty as shown in Table 3. For each candidate belief we calculate both the hypothetical action informed by the belief and the associated expected utility of the action under the belief (SEBU). For instance,

with belief b_1 in which the son is certainly innocent we see that the expected utility of proclaiming innocence is greater than that of proclaiming guilt. More concretely, $1 \times 3 + 0 \times 1 > 1 \times 0 + 0 \times 2$ and therefore a proclamation of innocence is chosen to inform the SEBU calculation of belief b_1. Once an action has been selected for a belief, the SEBU is simply the SEU of the selected action under the given belief. To continue our example, $SEBU\,(b_1) = 1 \times 3 + 0 \times 1 = 3$. Once the SEBU values for each of the candidate beliefs has been calculated, a subsequent maximization process designates the candidate belief with the maximal SEBU value as the desired belief which is b_1, the belief that the son is certainly innocent.

Table 1. Mother's Dilemma

	son innocent	son guilty
proclaim innocence	a	b
proclaim guilt	d	c

Table 2. Sample Utility Table for Mother

	son innocent	son guilty
proclaim innocence	3	1
proclaim guilt	0	2

Table 3. Candidate Belief Table

	$p\,(son\ innocent)$	$p\,(son\ guilty)$	action	SEBU
b_0	0.5	0.5	a_0	2.0
b_1	1.0	0.0	a_0	3.0
b_2	0.0	1.0	a_1	2.0

Generating the Candidate Belief Set. The candidate beliefs comprise the beliefs under *consideration* for the desired belief state. While in theory the candidate belief set may consist of any number of belief distributions, in practice however, the generation of candidate beliefs should be limited to a reasonable amount. Therefore, we only generate beliefs involving distributions of *certainty*.

For example, in a 3-state decision problem, 3 candidate beliefs will be generated as shown in Table 4.

Table 4. Candidate Beliefs in 3-State Decision Problem.

	$p(s_0)$	$p(s_1)$	$p(s_2)$
b_0	1	0	0
b_1	0	1	0
b_2	0	0	1

2.2 Belief Integration and Decision-Making

The purpose of the belief integration and decision-making phase is to choose an action while considering both the rational belief P and the desired belief P'. The manner in which this final decision is reached depends on both the type and magnitude of self-deception employed.

Mele distinguishes between two distinct forms of self-deception [15]:

- Being self-deceived into believing something that we desire to be true
- Being self-deceived into believing something we desire to be false

We call the former *optimistic* self-deception and the latter *pessimistic* self-deception.

Optimistic Self-Deception. The decision rule for optimistic self-deception is defined as follows in which $SEU(a)$ is the Subjective Expected Utility of action a, $SEU(P', a)$ is the Subjective Expected Utility of action a given the desired belief P', and α is a constant representing the magnitude of self-deception:

$$a_{\text{optimistic}} = \underset{a \in A}{\operatorname{argmax}} \left[(1 - \alpha) \times SEU(a) + \alpha \times SEU(P', a) \right] \tag{5}$$

Pessimistic Self-Deception. We characterize pessimistic self-deception as moving away from a desired belief state. Formally, we define it in a similar fashion to optimistic self-deception with the exception that the self-deceptive term is subtracted rather than added. The equation is as follows:

$$a_{\text{pessimistic}} = \underset{a \in A}{\operatorname{argmax}} \left[(1 - \alpha) \times SEU(a) - \alpha \times SEU(P', a) \right] \tag{6}$$

Magnitude of Self-Deception. Both optimistic and pessimistic definitions of self-deception utilize the constant α as a representation of the magnitude or strength of the self-deceptive tendencies evinced by a decision-maker. More formally, $0 \leq \alpha \leq 1$ and is defined such that when $\alpha = 0$ the decision-maker behaves in a purely rational manner as ascribed by SEU-Theory and when $\alpha = 1$ the decision-maker behaves in a purely self-deceptive manner in which all rational evidence is rejected and the desired belief is wholly adopted in either an optimistic or pessimistic fashion.

3 Simulation

Here we present our self-deceptive framework within the context of a game-theoretic simulation commonly referred to as the "Battle of the Sexes". With these experiments we seek to illustrate the behavior of both rational and self-deceptive agents as well as explore the interaction between the two.

The "Battle of the Sexes" traditionally represents a couple attempting to coordinate their actions for the evening without the benefit of communication. Their two choices are attending either an opera or a football game. Each partner has different preferences as to which event they'd like to attend. However, each partner would also rather attend their non-preferred event if it results in coordinating with their partner. At its core, the "Battle of the Sexes" is about coordination and synchronization since regardless of individual preferences, participants choosing to synchronize actions have higher utility both individually and collectively than they would alternatively. An illustrative utility matrix for the "Battle of the Sexes" is depicted in Table 5 in which the *row* player prefers attending a football game and the *column* player prefers the opera. Each entry in the table contains two utility values in which the first value refers to the utility received by the row player and the second value is the utility received by the column player.

Table 5. Example "Battle of the Sexes" payoff matrix

	Opera	Football
Opera	2,3	0,0
Football	1,1	3,2

3.1 Scenario Setup

In order to cast the "Battle of the Sexes" into a form amenable to analysis within our framework, we must probabilistically represent beliefs. Most traditional game-theoretic analyses focus on equilibrium strategies in which the

utilities for both participants is common knowledge. However, a probabilistic treatment of the game is appropriate in situations in which little or no information is available regarding a partner's preferences, strategies, or knowledge and when the only available information is probabilistic in nature, e.g., a relative frequency of past observations.

Consider the following scenario:

Example 2. Terry and Pat are players in the "Battle of the Sexes" in which Terry prefers attending the opera and Pat prefers football. We represent Terry's outcome preferences using the utilities shown in Table 6 which capture both Terry's primary goal of coordinating activities with Pat and a more general preference for opera. i

Let us assume that the initial beliefs for both players indicate an equally likely chance of attending either event. Irrespective of this rational belief, Terry's *desired* belief is one in which the possibility of Pat attending the opera is certain since this allows Terry to both coordinate events with Pat and attend the opera. Table 7 shows both Terry's rational and desired belief distributions. Figure 1 is a graph of Terry's decision thresholds with respect to the various decision-making processes described in this paper. A point on the graph is designated on the x-coordinate by α, representing the magnitude of self-deception, and on the y-coordinate by Terry's probabilistic estimate that Pat attends the football game. If the indicated point lies above the threshold curve of Terry's decision process Terry will choose to attend the football game. If it lies below the threshold curve Terry will attend the opera. For instance, when employing an optimistic self-deceptive decision-making process with $\alpha = .2$ and a rational belief that Pat's likelihood of attending the football game is .8, Terry will choose to attend the opera. However, given the same parameters utilizing a rational decision-making process, Terry will choose to attend the football game.

Table 6. Utility of outcomes for Terry

	Pat attends football game	Pat attends opera
Go to opera	1	3
Go to football game	2	0

3.2 Simulation Results and Discussion

We now present the experimental results for the scenario of Terry and Pat. In each of the six possible decision-making matchups we average the results over 500 runs in which each game is played successively for 200 rounds. Figure 2 depicts a graph of the mean utility per step for each agent in all of the six possible

Table 7. Belief distributions for Terry

	Pat attends football game p (football)	Pat attends opera p (opera)
Desired belief	0	1
Rational belief	.5	.5

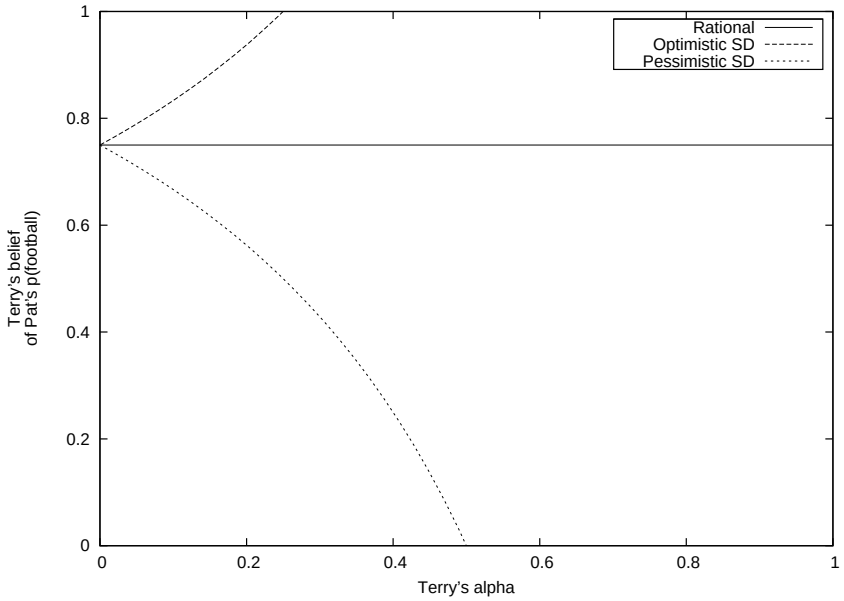

Fig. 1. Terry's Decision Threshold

matchup combinations. The graph in Fig. 3 shows the mean utility received in any given step for a particular matchup. Table 8 shows the approximate number of steps required in any given matchup to converge upon a stable solution of either coordination or miscoordination.

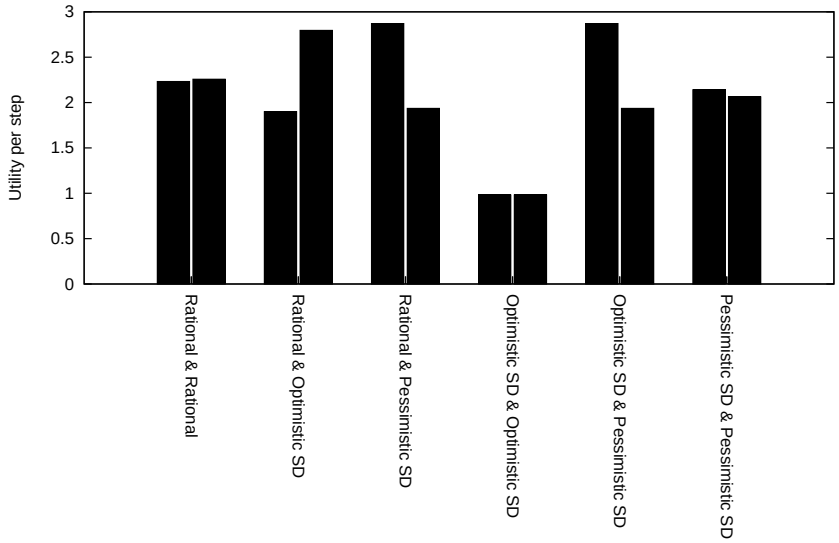

Fig. 2. Average Step Utility for "Battle of the Sexes"

Our experimental results show that situations involving participants employing dissimilar decision styles converge more quickly to a coordination of actions than do situations in which the participants employ identical decision styles. One situation in particular consisting of two agents employing a pessimistic self-deceptive style never attains a state of coordination while the other two combinations of identical decision processes take roughly 80 steps to reach coordination in contrast to the approximately 20 steps required for the combinations of dissimilar decision processes to converge on coordination.

Another interesting aspect of our experimental results is that in situations of eventual coordination, the agent that is most optimistic has a higher individual utility, i.e., always goes to its preferred event, than its partner. In situations where both participants utilize the same decision-making strategy, each partner is equally as likely to eventually emerge as the one attending its preferred event. Here we should note that in all cases of eventual coordination, once the first coordinating event is established, agents will continue coordinating on the same event for the duration of the game. For instance, the mean step utility of roughly 2.5 for two rational participants is an average of 500 runs and indicates the

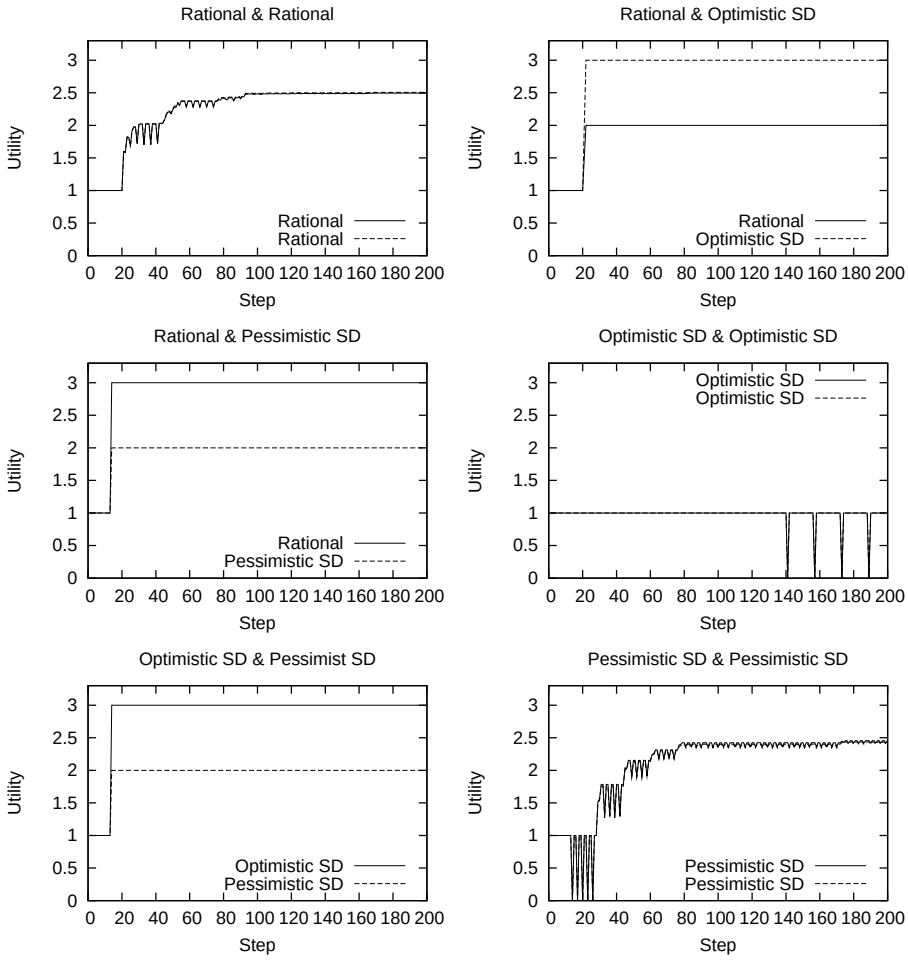

Fig. 3. Utility by step for "Battle of the Sexes"

Table 8. Convergence for "Battle of the Sexes"

Scenario	**Number of steps until convergence**	**Convergence Type**
Rational & Rational	80	Coordination
Rational & Optimistic SD	20	Coordination
Rational & Pessimistic SD	15	Coordination
Optimistic SD & Optimistic SD	0	Miscoordination
Optimistic SD & Pessimistic SD	15	Coordination
Pessimistic SD & Pessimistic SD	80	Coordination

equally likely possibility of the coordinating event being the preferred event of a given participant.

4 Related Work

In this paper we've attempted to operationalize the psychological concept of self-deception within a decision-theoretic framework. The notion of formalizing a psychological construct within the decision-theoretic domain is not without precedent. In fact, much of the work in decision theory since the groundbreaking efforts by Von Neumann, Morgenstern and Savage have concentrated on introducing a psychological dimension into the formal decision prcesses in order to provide for more robust and descriptive approaches. Regret Theory [12, 2] models the tendency to avoid decisions that could lead to an excessive feeling of regret. Prospect Theory [21] is a purely descriptive framework that employs a series of heursitics in order to approximate the mental shortcuts that people seem to employ when making decisions. Ellsberg's Index [7] and the Ambiguity Model of Einhorn and Hogarth [6] both model the perceived aversion to uncertainty that decision-makers sometimes express.

Within the realm of self-deception, Talbott presents a model based on the desirability of adopting some preferred belief. Talbott's notion of desirability is utility-based and is a weighting of the possiblity that the belief is accurate against the chance that it is not [20]. Based on this assessment, Talbott's model then calculates the expected utility for both behaving rationally and attempting to bias one's cognitive processes towards the desired belief. The primary difference between Talbott's work and the work presented in this paper is that Talbott defines the desirability of the possible belief outcomes externally while we derive that desirability using an agent's internal preferential utilities and then integrate the desired belief into the decision-making process using an externally defined constant representing the magnitude of deception.

In addition to decision-theoretic formulations of psychological phenomena, there exist a number of computational models of emotion and bias. The Affective Belief Revision system of Pimentel [17] describes a logic-based system of maintaining the consistency of a propositional knowledge base in which the belief revision activities are influenced by the affective state of the individual. Other computational models of emotion [9, 13] utilize self-deception as a coping mechanism to ease an agent's emotional stress. These computational models however, do not provide a manner in which to model the repercussions and tradeoffs of possibly adopting a false belief. The PsychSim modeling framework [18, 14], allows decision-theoretic agents to possibly influence the belief state of other agents by sending messages containing a hypothetical belief state. One factor that is assessed when evaluating these messages is self-interest. In other words, an agent will be more likely to accept a change in belief if the proposed belief is more amenable to the agent's desires and preferences. Since self-interest is evaluated entirely outside the reality of the current situtation, it is in principle similar to the notion of self-deception. The key difference between these computational

models of emotion and our work is that we present both the determination of the desired belief state and the subsequent process of self-deceptive belief revision all within a decision-theoretic framework.

5 Future Work

5.1 Alternative Decision Models

Future work may explore the implications of employing disparate decision models during the belief formulation phase and the subsequent belief integration and decision phase. For instance, a decision-maker may formulate the desired belief distribution based on SEU-Theory yet employ an alternative model such as Regret Theory or Prospect Theory in the actual decision-making phase.

5.2 Relaxed Formulation of Desired Belief

In this work, we choose the desired belief based upon the best possible outcome irrespective of the actual state distribution. However, in future work we may explore the possibility of a slightly altered and more relaxed definition of the desired belief. Specifically, rather than ignoring the reality of the situation in the formulation of the desired belief, we can choose a desired belief state *given* the current belief state. So once a course of action is chosen under the current belief state, we can then determine the outcome that is *desired*.

6 Conclusion

Whether the ultimate goal is to create more lifelike agents or simply model the actions of human decision-makers more accurately in order to make better decisions, understanding self-deception and exploring the computational aspect of the phenomena is key. In this work we've detailed a descriptive framework for modeling the psychological act of self-deception within a decision-theoretic environment based on the tenets of SEU-Theory. Our self-deceptive theory utilizes SEU-Theory for not only the desired belief integration and decision-making process but also for the formation of the desired belief state that is central to the biasing processes of self-deception. Through a series of experimental simulations using the "Battle of the Sexes" game-theoretic formulation we've shown that our framework operationalizes both optimistic and pessimistic self-deception processes and that within certain situations, a healthy dose of self-deception is beneficial.

References

1. E. Aronson. Dissonance Theory: Progress and Problems. *Contemporary Issues in Social Psychology*, 2:310–323, 1968.

2. D.E. Bell. Regret in Decision Making under Uncertainty. *Operations Research*, 30(5):961–981, 1982.
3. T. Bickmore, A. Gruber, and R. Picard. Establishing the Computer–Patient Working Alliance in Automated Health Behavior Change Interventions. *Patient Education and Counseling*, 59(1):21–30, 2005.
4. R.B. Cialdini. *Influence: Science and Practice*. Allyn and Bacon, 2001.
5. R. Demos. Lying to Oneself. *The Journal of Philosophy*, 57(18):588–595, 1960.
6. H.J. Einhorn and R.M. Hogarth. Decision Making Under Ambiguity. *The Journal of Business*, 59(4):225–250, 1986.
7. D. Ellsberg. Risk, Ambiguity, and the Savage Axioms. *Rationality in Action: Contemporary Approaches*, 1990.
8. L. Festinger. A Theory of Cognitive Dissonance. *Evanston, IL: Row, Peterson*, 1(958):65–86, 1957.
9. J. Gratch and S.C. Marsella. A Domain-Independent Framework for Modeling Emotion. *Cognitive Systems Research*, 5(4):269–306, 2004.
10. Z. Kunda. The Case for Motivated Reasoning. *Psychological Bulletin*, 108(3):480–498, 1990.
11. R.S. Lazarus. *Emotion and Adaptation*. Oxford University Press, USA, 1991.
12. G. Loomes and R. Sugden. Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty. *The Economic Journal*, 92(368):805–824, 1982.
13. S.C. Marsella and J. Gratch. Modeling Coping Behavior in Virtual Humans: Don't Worry, be Happy. *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 313–320, 2003.
14. S.C. Marsella, D.V. Pynadath, and S.J. Read. PsychSim: Agent-Based Modeling of Social Interactions and Influence. *Proceedings of the International Conference on Cognitive Modeling*, pages 243–248, 2004.
15. A.R. Mele. Understanding and Explaining Real Self-Deception. *Behavioral and Brain Sciences*, 20(01):127–134, 1997.
16. A.R. Mele. *Self-Deception Unmasked*. Princeton University Press, 2000.
17. C.F. Pimentel and M.R. Gravo. Affective Revision. *Progress in Artificial Intelligence: 12th Portuguese Conference on Artificial Intelligence, EPIA 2005, Covilhã, Portugal, December 5-8, 2005: Proceedings*, 2005.
18. D.V. Pynadath and S.C. Marsella. PsychSim: Modeling Theory of Mind with Decision-Theoretic Agents. *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1181–1186, 2005.
19. L.J. Savage. *The Foundations of Statistics*. Courier Dover Publications, 1972.
20. W.J. Talbott. Intentional Self-Deception in a Single Coherent Self. *Philosophy and Phenomenological Research*, 55(1):27–74, 1995.
21. A. Tversky and D. Kahneman. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263–292, 1979.