Self-Deceptive Decision Making: Normative and Descriptive Insights

Jonathan Y. Ito, David V. Pynadath, Stacy C. Marsella
USC Institute for Creative Technologies
13274 Fiji Way
Marina del Rey, CA 90292
{ito,pynadath,marsella}@ict.usc.edu

ABSTRACT

Computational modeling of human belief maintenance and decision-making processes has become increasingly important for a wide range of applications. We present a framework for modeling the psychological phenomenon of self-deception in a decision-theoretic framework. Specifically, we model the self-deceptive behavior of wishful thinking as a psychological bias towards the belief in a particularly desirable situation or state. By leveraging the structures and axioms of Expected Utility (EU) Theory we are able to operationalize both the determination and the application of the desired belief state with respect to the decision-making process of expected utility maximization. While we categorize our framework as a descriptive model of human decision making, we show that in certain circumstances the realized expected utility of an action biased by wishful thinking can exceed that of an action motivated purely by the maximization of perceived expected utility. Finally, we show that our framework of self-deception and wishful thinking has the descriptive flexibility to account for the inconsistencies highlighted by the Common Ratio Effect and the Allais Paradox.

Categories and Subject Descriptors

I.2.0 [**Artificial Intelligence**]: General—*cognitive simulation*; I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*multiagent systems*; I.2.3 [**Artificial Intelligence**]: Deduction and Theorem Proving—*nonmonotonic reasoning and belief revision*; J.4 [**Social and Behavioral Sciences**]: Psychology

General Terms

Algorithms, Human Factors, Theory

Keywords

Self-deception, wishful thinking, decision theory, multi-agent systems

1. INTRODUCTION

Psychological bias is an unavoidable factor when human decision makers are faced with complex decisions in an uncertain environment. Our beliefs are not formed merely by the evidence at hand but are influenced by our desires and intentions. Research on human behavior has identified a range of rational as well as seemingly irrational tendencies in how people manage their beliefs and make decisions [12]. 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 [15]. Research on cognitive dissonance [8] has demonstrated that people often seek to achieve consistency between their beliefs and behaviors and focuses on how we alter beliefs in order to resolve inconsistencies between a desired positive self-image and our behavior [2]. 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 [13].

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 [7] can be computationally modeled [17] for a variety of applications such as health interventions designed to alter user behavior [6].

In previous work we introduced the computational notion of wishful thinking and showed that in some iterated games wishful thinking outperforms perceived expected utility maximization [10]. In this work, we formalize a general framework of self-deception and 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 Expected Utility Theory. While embedding the seemingly irrational process of wishful thinking and self-deception into a decision-theoretic framework may in itself seem irrational, we contend that seemingly irrational behavior such as wishful thinking, motivated inference, and self-deception can be grounded and integrated with an agent's expected utility calculations in a principled fashion.

Specifically, we model the self-deceptive behavior of wishful thinking as a psychological bias towards the belief in a particularly desirable situation or state. By leveraging

Cite as: Self-Deceptive Decision Making: Normative and Descriptive Insights, Jonathan Y. Ito, David V. Pynadath, Stacy C. Marsella, *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009),* Decker, Sichman, Sierra and Castelfranchi (eds.), May, 10–15, 2009, Budapest, Hungary, pp. 1113–1120

	s_1	s_2	$\dots$	s_j	$\dots$	s_n
A_1	x_{11}	x_{12}	$\dots$	x_{1j}	$\dots$	x_{1n}
A_2	x_{21}	x_{22}	$\dots$	x_{2j}	$\dots$	x_{2n}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
A_i	x_{i1}	x_{i2}	$\dots$	x_{ij}	$\dots$	x_{in}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
A_m	x_{m1}	x_{m2}	$\dots$	x_{mj}	$\dots$	x_{mn}

Table 1: Generalized Decision Problem

the structures and axioms of Expected Utility (EU) Theory [19] and Subjective Expected Utility (SEU) Theory [18] we are able to operationalize both the determination and the application of the deceptive belief state motivated by wishful thinking with respect to the decision-making process of expected-utility maximization. While our theory of self-deception is motivated by a psychological characterization of human decision making, we also show that the theory has important normative implications as well. In particular, we show that in some situations of uncertainty in which particular errors, e.g., the evaluation of preference, incorrect assumptions of causal structure, are present, the realized expected utility of an action biased by wishful thinking can exceed that of an action motivated purely by the maximization of expected utility. Finally, in order to characterize the descriptive flexibility of our framework, we show that our wishful-thinking formulation can account for the effects seen in both the Common Ratio Effect and the Allais Paradox, two well-documented instances of human behavior inconsistent with EU Theory.

2. SELF-DECEPTION

We define a general decision problem as a choice between m distinct actions in an environment consisting of n possible states. A specific outcome x_{ij} is obtained by performing action A_i when the prevailing state of nature is s_j as depicted in Table 1. The probability of being in a given state of nature s_j is given by the probability mass function $pr\left(s_j\right)$. In decision problems involving some degree of uncertainty, the true probability distribution is unknown and therefore $pr\left(s_j\right)$ is considered the subjective probability estimate of state s_j by the decision maker. Therefore, the probability distribution function, pr, is considered representative of the *beliefs* of the decision maker.

To arrive at an operationalized definition of self-deceptive decision making we will define what a self-deceptive belief is, the manner in which it is integrated with the rational, unbiased beliefs of a decision maker, and how the decision process is altered by the inclusion of self-deception. Furthermore, we will also define wishful thinking as a specific instantiation of self-deception in which we further specify the process by which the self-deceptive belief is designated.

2.1 Self-Deceptive Beliefs

Self-deception is the belief in P when the totality of available evidence suggests $\neg P$. Computationally we define a self-deceptive belief as an *alternate* probability distribution $pr^*\left(s_j\right)$, which we refer to as the *deceptive belief state*, that is distinct from the probability distribution $pr\left(s_j\right)$. This deceptive belief state may be an optimistic belief that the best possible outcome will occur, as we later describe in our formulation of wishful thinking, or even a pessimistic belief that the worst possible outcome will be realized.

Psychological research within the realm of motivated reasoning asserts that the reasoning processes of individuals are influenced by both the desire to believe what we want and the desire to be as accurate and rational as possible [12]. In a similar manner we define the *act* of self-deception as a mediation between the deceptive belief state, $pr^*\left(s_j\right)$, and the unbiased belief state, $pr\left(s_j\right)$, resulting in the compound probability distribution $pr_{sd}\left(s_j\right)$. Furthermore, for computational and conceptual simplicity we impose the restriction that $pr_{sd}\left(s_j\right)$ is a linear combination of the deceptive and unbiased belief states as in (1).

$$pr_{sd}\left(s_j\right) = (1-\alpha)\, pr\left(s_j\right) + \alpha pr^*\left(s_j\right) \tag{1}$$

By varying the value of α, which we refer to as the self-deceptive constant, the degree of self-deception evinced by the decision maker can be tuned such that when $\alpha = 0$ the decision maker is fully rational and when $\alpha = 1$ the decision maker is fully delusional.

2.2 Self-Deceptive Decision Making

Our self-deceptive implementation of decision making is based upon EU Theory and utilizes the deceptive beliefs formulated in the previous section. According to EU Theory, we may define a utility function $\mu\left(x_{ij}\right)$ representing the preferences of a decision maker over the possible set of outcomes. Furthermore, EU Theory states that by selecting action A_u among the available set of actions such that expected utility is maximized as in (2), the decision maker is acting in accord with its desires.

$$u = \operatorname*{argmax}_{i=1}^{m} \sum_{j=1}^{n} pr\left(s_j\right) \cdot \mu\left(x_{ij}\right) \tag{2}$$

In order to operationalize self-deceptive decision making we restate (2) and replace the unbiased belief, $pr\left(s_j\right)$, with the psychologically biased belief, $pr_{sd}\left(s_j\right)$, to arrive at the self-deceptive selection of Action A_{sd} as in (3).

$$\begin{aligned} sd &= \operatorname*{argmax}_{i=1}^{m} \sum_{j=1}^{n} pr_{sd}\left(s_j\right) \cdot \mu\left(x_{ij}\right) \\ &= \operatorname*{argmax}_{i=1}^{m} \sum_{j=1}^{n} \left((1-\alpha)\, pr\left(s_j\right) + \alpha pr^*\left(s_j\right)\right) \cdot \mu\left(x_{ij}\right) \end{aligned} \tag{3}$$

2.3 Wishful Thinking

We define wishful thinking as a specific instance of self-deception in which a decision maker is biased towards believing that positive outcomes are more likely to occur than reality would suggest. Computationally, we define the desired belief state specified through wishful thinking as the probability distribution $pr_w\left(s_j\right)$ which would maximize the expected utility of the decision maker. It follows that this is the probability distribution in which the state of nature, s_c, that is required for the most preferred outcome is certain, i.e., assigned a probability of 1. In other words, a belief based on wishful thinking is the belief that if true, would result in the most preferred outcome. The determination of c is operationalized in (4).

$$c = \operatorname*{argmax}_{j=1}^{n} \left(\max_{i=1}^{m} \mu(x_{ij}) \right) \tag{4}$$

This particular formulation of self-deception allows us to concretely specify the desired belief state by leveraging the preferences of the decision maker encoded within EU Theory. We can now state the wishful-thinking based selection of action A_w as (5).

$$w = \operatorname*{argmax}_{i=1}^{m} (1-\alpha) \sum_{j=1}^{n} pr(s_j) \cdot \mu(x_{ij}) + \alpha\mu(x_{ic}) \tag{5}$$

A noteworthy property of wishful-thinking based decision making is that when $\alpha = 0$, i.e., the degree of self-deception evinced by the decision maker is very low, the equation collapses into a traditional EU-maximization process. Inversely, when the degree of self-deception is very high the decision maker disregards the unbiased probability distribution and bases its decision wholly on deceptive beliefs.

3. NORMATIVE POSSIBILITY OF WISHFUL THINKING

The axioms supporting EU Theory provide a clean and elegant definition of rationality that is easily amenable to both implementation and analysis. In practice however, the strategy of EU maximization may not always yield optimal results. In this section, we provide a general analysis of the conditions under which wishful thinking may be preferable to EU maximization. In particular, we characterize certain types of plausible errors in judgment which may be present in the decision-making process and establish the conditions necessary such that reasoning biased by wishful thinking outperforms reasoning motivated solely by expected utility maximization.

3.1 Errors of Causality: Illusions of Control and Reduced Estimates of Control

Langer defines the illusion of control as an "expectancy of a personal success probability inappropriately higher than the objective probability would warrant" [14]. Research has shown that people actively engaged in a decision or task often perceive a greater amount of control than actually exists [3, 5]. For instance, a shooter in the game of craps may feel justified making a large wager on his or her next roll of the dice under the illusion that active participation will engender a favorable outcome. Inversely, research has shown that negative moods associated with depression are often associated with reduced estimates of control [1] in which people engaged in a task perceive *less* control than actually exists. In these situations, it may be possible to offset reduced estimations of control through the application of self-deceptive wishful thinking in decision-support systems.

Assume that the actual probability model, $pr(s_j / A_i)$, is one of positive control in which performing action A_i increases the likelihood of the state leading to the most preferred outcome for action A_i by a factor of t as seen in (6) where $0 \leq t \leq 1$.

$$pr(s_j / A_i) = \begin{cases} (1-t)\,pr(s_j) + t & \text{if } j = \operatorname{argmax}_{k=1}^{n} \mu(x_{ik}) \\ (1-t)\,pr(s_j) & \text{if } j \neq \operatorname{argmax}_{k=1}^{n} \mu(x_{ik}) \end{cases} \tag{6}$$

We then state the necessary conditions under which the realized expected utility for the self-deceptive choice of A_w exceeds that of A_u, which maximizes perceived expected utility, as (7) in which x_{ua} and x_{wb} are the most preferable outcomes when taking actions A_u and A_w respectively.

$$\begin{aligned} &\text{if } (1-t) \sum_{j=1}^{n} pr(s_j) \left(\mu(x_{wj}) - \mu(x_{uj}) \right) > \\ &\quad t \left(\mu(x_{ua}) - \mu(x_{wb}) \right), \text{ then } A_w \text{ outperforms } A_u \end{aligned} \tag{7}$$

In general, when reduced estimations of control are minimal, then the act which maximizes expected utility outperforms the self-deceptive act. However, when the decision maker significantly underestimates its ability at affecting a positive outcome then the self-deceptive act may offset this and outperform the action maximizing expected utility.

3.2 Errors in Utility Assessment: Regret and Rejoicing

In addition to errors in subjective probability assessment, decision makers may experience difficulty expressing their preferences completely, consistently, and unambiguously. Furthermore, once an outcome has been obtained there may be additional psychological biases affecting realized utility such as feelings of regret or rejoicing which are described in Regret Theory [16, 4]. The theory states that the amount of regret or rejoicing that is experienced upon obtaining a particular outcome is measured in relation to the outcome had a different action been chosen and that these psychological factors play critical roles in human decision making.

Regret Theory contends that decision makers utilize a modified utility function, $M(x_{ij}, x_{kj})$ as in (8) in which x_{ij} is the realized outcome and x_{kj} is the outcome which would have been obtained had another action been chosen. The modified utility function of Regret Theory is based on two distinct measures: A choiceless utility function and a regret-rejoicing component. The choiceless utility function $C(x_{ij})$ describes the utility of outcome x_{ij} when the agent has no active participation in the decision-making process and is often likened to the general notion of utility employed in many decision problems. The regret-rejoicing function $R(\cdot)$ is strictly increasing and three times differentiable.

$$M(x_{ij}, x_{kj}) = C(x_{ij}) + R\left(C(x_{ij}) - C(x_{kj})\right) \tag{8}$$

Assuming that the modified utility function, $M(x_{ij}, x_{kj})$, as opposed to the choiceless interpretation of utility, is consistent with the preferences of a decision maker and that the same decision maker erroneously chooses an action based on the standard, or choiceless, interpretation of utility we can then state the conditions under which the realized expected utility for the self-deceptive choice of A_w exceeds that of A_u, which maximizes expected utility, as in (9).

$$\begin{aligned} &\text{if } \sum_{j=1}^{n} pr(s_j) \left(C(x_{wj}) + R\left(C(x_{wj}) - C(x_{uj})\right) \right) > \\ &\quad \sum_{j=1}^{n} pr(s_j) \left(C(x_{uj}) + R\left(C(x_{uj}) - C(x_{wj})\right) \right), \\ &\text{then } A_w \text{ outperforms } A_u \end{aligned} \tag{9}$$

	B_1	...	B_j	...	B_n
A_1	x_{11}, y_{11}	...	x_{1j}, y_{j1}	...	x_{1n}, y_{n1}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
A_i	x_{i1}, y_{1i}	...	x_{ij}, y_{ji}	...	x_{in}, y_{ni}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
A_m	x_{m1}, y_{1m}	...	x_{mj}, y_{jm}	...	x_{mn}, y_{nm}

Table 2: 2-Player Game

3.3 Errors of Environment: Competitive Decision Making

Thus far our discussion has been restricted to individual decision making in environments of risk and uncertainty, i.e., decisions against nature, in which nature is characterized as an uncaring opponent having no regard for the preferences of the decision maker. Within this type of environment, in which the likelihood of the various states may be evaluated, our model of self-deception is well-suited. However, many decisions can only be categorized as competitive or cooperative situations, otherwise known as games, in which the outcome depends on the actions of two or more participants each possessing their own, possibly unique, set of preferences.

A typical representation of a two-player game is given in Table 2 in which player A has a choice of m actions and player B chooses between n actions. The outcome for any given pair of actions A_i and B_j is given as x_{ij} for player A and y_{ji} for player B.

In the context of analyzing games, making probabilistic evaluations as to the action, B_k, of player B is difficult since the action determination of B may rely largely upon its beliefs regarding the intentions of player A. Thus, for the purpose of our analysis, it is meaningful to categorize player B according to the epistemic knowledge it possesses since it is this knowledge which ultimately informs its strategy in determining an action.

In the following sections we categorize player B as either assuming a static environment, assuming player A employs EU maximization, or assuming A is engaged in wishful-thinking. These categorizations of player B will enable us to further define the method by which player B comes to an action determination which in turn allows us to perform a more thorough analysis on the realized utility for player A.

3.3.1 *Player* B *Assumes Static Environment*

The assumption of a static decision environment implies that the decision maker believes the environment can be characterized as a game against nature and thus probabilistic judgments over the various states are made. If player B is operating under this assumption and ascribes to the principles of EU Theory then its selection of action B_k rests solely on its preferences over outcomes and its probabilistic assessment of states such that action B_k is chosen as in (10) where $pr_B(A_i)$ is player B's subjective evaluation of the likelihood A_i, which in actuality is the action chosen by player A, will occur.

$$k = \operatorname*{argmax}_{j=1}^{n} \sum_{i=1}^{m} pr_B(A_i) \cdot \mu(y_{ji}) \tag{10}$$

Given player B's utility function and probability distribution, its action determination becomes deterministic. Therefore, to characterize the situations in which the self-deceptive choice of A_w outperforms the EU-maximization choice of A_u for player A when player B assumes a static decision environment we have (11).

$$\begin{aligned}
&\text{if } \mu(x_{wk}) > \mu(x_{uk}), \text{ then } A_w \text{ outperforms } A_u \text{ where} \qquad (11)\\
&k = \operatorname*{argmax}_{j=1}^{n} \sum_{i=1}^{m} pr_B(A_i) \cdot \mu(y_{ji})\\
&w = \operatorname*{argmax}_{i=1}^{m} (1-\alpha) \sum_{j=1}^{n} pr_A(B_j) \cdot \mu(x_{ij}) + \alpha\mu(x_{ic})\\
&c = \operatorname*{argmax}_{j=1}^{n} \left(\max_{i=1}^{m} \mu(x_{ij}) \right)\\
&u = \operatorname*{argmax}_{i}^{m} \sum_{j}^{n} pr_A(B_j) \cdot \mu(x_{ij})
\end{aligned}$$

If agent B chooses the action B_k that is coincident with the self-deceptive belief of player A as defined in accordance to our wishful-thinking formulation we see that the realized utility for the self-deceptive choice of A_w will meet or exceed that of A_u, which maximizes expected utility, as in (12).

$$\begin{aligned}
&\text{if } \operatorname*{argmax}_{j=1}^{n} \sum_{i=1}^{m} pr_B(A_i) \cdot \mu(y_{ji}) = \operatorname*{argmax}_{j=1}^{n} \left(\max_{i=1}^{m} \mu(x_{ij}) \right),\\
&\text{then } \mu(x_{wk}) \geq \mu(x_{uk}) \qquad (12)
\end{aligned}$$

In situations where player B assumes a static decision environment, determining if A's choice of a self-deceptive action exceeds the realized expected utility of an action chosen strictly through perceived expected utility maximization for player A depends on the beliefs of both player A and player B as well as the payout structure of the game since the choices of both players are independent of each other.

3.3.2 *Player* B *Assumes Opponent is an EU Maximizer*

When player B knows it is operating within a competitive decision-making environment, its choice of action B_k will depend largely on its beliefs regarding the intentions of player A. If B believes that A will behave as if maximizing expected utility by choosing action A_u, then B will do best by choosing action B_k such as to maximize its own utility under the expectation that A will choose A_u as in (13).

$$k = \operatorname*{argmax}_{j=1}^{n} \mu(y_{ju}) \tag{13}$$

Given player B's utility function and probability distribution, its action determination becomes deterministic. Therefore, to characterize the situations in which the self-deceptive choice of A_w outperforms the EU-maximization choice of A_u for player A when player B assumes an EU-maximization opponent we have (14)

$$\text{if } \mu(x_{wk}) > \mu(x_{uk}), \text{ then } A_w \text{ outperforms } A_u \text{ where} \quad (14)$$
$$k = \operatorname*{argmax}_{j=1}^{n} \mu(y_{ju})$$
$$w = \operatorname*{argmax}_{i=1}^{m} (1-\alpha)\sum_{j=1}^{n} pr_A(B_j)\cdot\mu(x_{ij}) + \alpha\mu(x_{ic})$$
$$c = \operatorname*{argmax}_{j=1}^{n} \left(\max_{i=1}^{m}\mu(x_{ij})\right)$$
$$u = \operatorname*{argmax}_{i=1}^{m} \sum_{j=1}^{n} pr_A(B_j)\cdot\mu(x_{ij})$$

If agent B chooses the action B_k that is coincident with the self-deceptive belief of player A as defined in accordance to our wishful-thinking formulation we see that the realized utility for the self-deceptive choice of A_w will meet or exceed that of A_u, which maximizes expected utility, as in (15).

$$\text{if } \operatorname*{argmax}_{j=1}^{n}\mu(y_{ju}) = \operatorname*{argmax}_{j=1}^{n}\left(\max_{i=1}^{m}\mu(x_{ij})\right),$$
$$\text{then } \mu(x_{wk}) \geq \mu(x_{uk}) \quad (15)$$

3.3.3 Player B Assumes Opponent Engages in Wishful Thinking

When player B is aware of both the game-theoretic environment under which it operates and the self-deceptive decision process employed by its opponent, it will correctly expect A to choose the self-deceptive action A_w and will choose an action B_k which maximizes its own utility as in (16).

$$k = \operatorname*{argmax}_{j=1}^{n}\mu(y_{jw}) \quad (16)$$

Given player B's utility function and probability distribution, its action determination becomes deterministic. Therefore, to characterize the situations in which the self-deceptive choice of A_w outperforms the EU-maximization choice of A_u for player A when player B assumes a self-deceptive opponent we have (17).

$$\text{if } \mu(x_{wk}) > \mu(x_{uk}), \text{ then } A_w \text{ outperforms } A_u \text{ where} \quad (17)$$
$$k = \operatorname*{argmax}_{j=1}^{n} \mu(y_{jw})$$
$$w = \operatorname*{argmax}_{i=1}^{m} (1-\alpha)\sum_{j=1}^{n} pr_A(B_j)\cdot\mu(x_{ij}) + \alpha\mu(x_{ic})$$
$$c = \operatorname*{argmax}_{j=1}^{n} \left(\max_{i=1}^{m}\mu(x_{ij})\right)$$
$$u = \operatorname*{argmax}_{i=1}^{m} \sum_{j=1}^{n} pr_A(B_j)\cdot\mu(x_{ij})$$

If agent B chooses the action B_k that is coincident with the self-deceptive belief of player A as defined in accordance to our wishful-thinking formulation we see that the realized utility for the self-deceptive choice of A_w will meet or exceed that of A_u, which maximizes expected utility, as in (18).

$$\text{if } \operatorname*{argmax}_{j=1}^{n}\mu(y_{jw}) = \operatorname*{argmax}_{j=1}^{n}\left(\max_{i=1}^{m}\mu(x_{ij})\right),$$
$$\text{then } \mu(x_{wk}) \geq \mu(x_{uk}) \quad (18)$$

Note that in zero-sum games in which the preferences of player A are directly opposed by those of player B, such that $\mu(x_{ij}) + \mu(y_{ji}) = 0$, player B will never choose the action B_k that is the preferred belief state of player A. For an abbreviated proof of this situation, we may represent the action determination of B in a zero-sum game as (19). Assume to the contrary that B_k is the desired belief of A. Then according to our wishful-thinking formulation, $\mu(x_{ik}) > \mu(x_{ij})$. We also know from (19) that $-\mu(x_{ik}) > -\mu(x_{ij})$ or $\mu(x_{ik}) < \mu(x_{ik})$ which is a contradiction.

$$k = \operatorname*{argmax}_{B_j}^{B_n} -\mu(x_{wj}) \quad (19)$$

In games in which player B correctly assumes that its opponent is engaged in self-deception, whether player A's choice of a self-deceptive action outperforms the action chosen to maximize perceived expected utility depends entirely on the payoff structure of the game and the degree of self-deception evinced by player A.

4. DESCRIPTIVE CHARACTERIZATION OF SELF-DECEPTION

Many of the alternatives to EU Theory such as Prospect Theory [11] and Regret Theory [4, 16] tend toward descriptive characterizations of decision making aimed at resolving observational inconsistencies in EU Theory. Here we present how our model of self-deception accounts for the findings in two perceived paradoxes in EU Theory to demonstrate the descriptive flexibility of our framework. Specifically, we will show that our wishful-thinking formulation of self-deception can account for the effects of both the Certainty Effect and the Allais Paradox.

In their development of Prospect Theory, Kahneman and Tversky collected a series of preference observations between pairs of gambles or prospects [11], a subset of which are shown in Table 3. The paradoxical nature of these preference orderings arises when one attempts to assign consistent utility values for the various payouts presented in each paired set of decision problems such that the stated preference orderings are maintained. According to these preference orderings there exist no valid and consistent utility mappings for the various payouts and therefore EU Theory is unable to account for these statistically significant observations without the admission of additional dimensions of utility.

Here we will attempt to provide a meaningful characterization of the descriptive facilities of our self-deceptive framework by stating the conditions necessary to account for the results seen in the Certainty Effect and Allais Paradox. In particular, we show that under certain circumstances we can establish a valid utility mapping for the payouts and assign a valid value for α, representing the magnitude of the self-deception, such that $0 \leq \alpha \leq 1$.

Thus, through observing the behavior of a human decision maker we wish to model (or by authoring the behavior of a software agent), we can collect a sample of representative decisions as in Table 3 such that our results in this section allow us to translate each decision into a constraint on possible α values, i.e., degree of self-deception exhibited. By analyzing the resulting constraints, we can isolate situations where decision makers show consistent self-deception across

Prospects offered	Preference	Paradox
$X_1 = (4000, 0.80)$ $X_2 = (3000, 1.00)$	$X_1 \prec X_2$	Certainty Effect
$X_3 = (4000, 0.20)$ $X_4 = (3000, 0.25)$	$X_3 \succ X_4$	
$X_5 = (2500, 0.33; 2400, 0.66)$ $X_6 = (2400, 1.00)$	$X_5 \prec X_6$	Allais Paradox
$X_7 = (2500, 0.33)$ $X_8 = (2400, 0.34)$	$X_7 \succ X_8$	

Table 3: Prospects and Preferences

\$	μ (\$)
4000	1
3000	k
0	0

Table 4: Certainty Effect: Utility Assignment

multiple problems, as well as the situations where even their self-deception is inconsistent. This precise characterization will better enable us to identify important modeling features we need to add in order to refine our self-deception model.

4.1 Certainty Effect

The Certainty Effect or Common Ratio Effect is the phenomenon in which outcomes which are certain, i.e., probabilities approaching unity, are more highly preferred than conventional EU Theory would suggest. The certainty effect is evidenced in the conjunction of the preferences $X_1 \prec X_2$ and $X_3 \succ X_4$ of Table 3. This conjunction of preferences is contrary to EU Theory in that there does not exist any valid mapping of utility values to the payoffs such that the designated preference orderings hold.

Since utility is defined up to a positive linear transformation, we may represent the utility of the various monetary outcomes as shown in Table 4 such that $0 < k < 1$. Assuming that the choice between two prospects are independent, we can represent the choice between X_1 and X_2 as a standard decision problem consisting of the states depicted in Table 5 and outcomes in Table 6.

According to our wishful-thinking formulation, certainty in either a_1 *or* a_2 is acceptable as a valid deceptive belief state. So we represent the self-deceptive preference of $X_1 \prec X_2$ as a constraint on the value of α as in (20), which depends both on the selection of k and the determination of the preferred belief state.

state	X_1	X_2	$pr(a_j)$
a_1	4000	3000	0.80
a_2	4000	0	0.00
a_3	0	3000	0.20
a_4	0	0	0.00

Table 5: Certainty Effect: State Distribution for X_1,X_2

	a_1	a_2	a_3	a_4
X_1	1	1	0	0
X_2	k	0	k	0

Table 6: Certainty Effect: Outcomes for X_1,X_2

state	X_3	X_4	$pr(b_j)$
b_1	4000	3000	0.05
b_2	4000	0	0.15
b_3	0	3000	0.20
b_4	0	0	0.60

Table 7: Certainty Effect: State Distribution for X_3,X_4

$$\alpha < \begin{cases} 5k-4 & \text{if } a_1 \text{ is preferred} \\ \frac{5k-4}{1+5k} & \text{if } a_2 \text{ is preferred} \end{cases} \tag{20}$$

Similarly we represent the decision between X_3 and X_4 as a standard decision problem consisting of the states depicted in Table 7 and outcomes in Table 4.1.

According to the wishful-thinking formulation, certainty in either b_1 *or* b_2 is acceptable as a valid deceptive belief state. So we can represent the self-deceptive preference of $X_3 \succ X_4$ as a constraint on the value of α as in (21), which depends both on the selection of k and the determination of the preferred belief state.

$$\alpha > \begin{cases} \frac{5k-4}{16-15k} & \text{if } b_1 \text{ is desired} \\ \frac{5k-4}{16+5k} & \text{if } b_2 \text{ is desired} \end{cases} \tag{21}$$

The final determination of whether a valid mapping for both α and k exists for the Certainty Effect depends on the combination of preferred belief states chosen by the self-deceptive agent. The valid space of mappings for α and k for all 4 combinations of preferred belief states are shown in Figure 1. Notice that when the desired belief states are a_2 and b_1, no valid mapping exists.

4.2 Allais Paradox

The Allais paradox is an inconsistency observed in EU Theory in which the Sure-Thing Principle is violated as seen in the conjunction of preferences $X_5 \prec X_6$ and $X_7 \succ X_8$ shown in Table 3. This conjunction of preferences is contrary to EU Theory in that there does not exist any valid mapping of utility values to the payoffs such that the designated preference orderings hold.

	b_1	b_2	b_3	b_4
X_3	1	1	0	0
X_4	k	0	k	0

Table 8: Certainty Effect: Outcomes for X_3,X_4

Since the utility function is defined up to a positive linear transformation, we may represent the utility of the various monetary outcomes as shown in Table 9 such that $0 < k < 1$. Assuming that the choice between two prospects are inde-

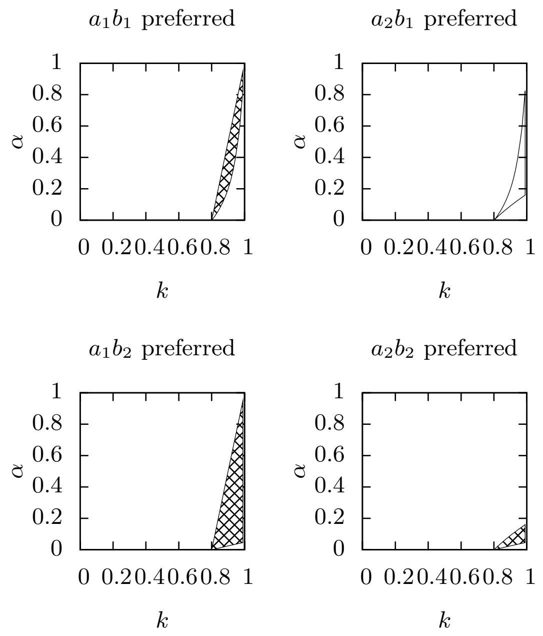

Figure 1: Certainty Effect: Valid α and k Mappings

\$	μ (\$)
2500	1
2400	k
0	0

Table 9: Allais Paradox: Utility Assignment

pendent, the choice between X_5 and X_6 is depicted as a standard decision problem consisting of the states shown in Table 10 and the outcomes in Table 11.

According to the wishful-thinking formulation, certainty of either a_1 *or* a_2 are both valid preferred belief states so we can represent the self-deceptive preference of $X_5 \prec X_6$ as a constraint on the value of α as in (22), which depends both on the selection of k and the determination of the preferred belief state.

$$\alpha < \begin{cases} \frac{34k-33}{67-66k} & \text{if } a_1 \text{ is preferred} \\ \frac{34k-33}{67+34k} & \text{if } a_2 \text{ is preferred} \end{cases} \tag{22}$$

state	X_5	X_6	$pr(a_j)$
a_1	2500	2400	0.33
a_2	2500	0	0.00
a_3	2400	2400	0.66
a_4	2400	0	0.00
a_5	0	2400	0.01
a_6	0	0	0.00

Table 10: Allais Paradox: State Distribution for X_5, X_6

	a_1	a_2	a_3	a_4	a_5	a_6
X_5	1	1	k	k	0	0
X_6	k	0	k	0	k	0

Table 11: Allais Paradox: Outcomes for X_5, X_6

state	X_7	X_8	$pr(b_j)$
b_1	2500	2400	0.1122
b_2	2500	0	0.2178
b_3	0	2400	0.2278
b_4	0	0	0.4422

Table 12: Allais Paradox: State Distribution for X_7, X_8

Similarly we represent the decision between X_7 and X_8 as a standard decision problem consisting of the states depicted in Table 12 and the outcomes in Table 13.

According to the wishful-thinking formulation, certainty of either b_1 *or* b_2 are both valid preferred belief states and so we can represent the self-deceptive preference of $X_7 \succ X_8$ as a constraint on the value of α as in (23), which depends both on the selection of k and the determination of the preferred belief state.

$$\alpha > \begin{cases} \frac{34k-33}{67-66k} & \text{if } b_1 \text{ is desired} \\ \frac{34k-33}{67+34k} & \text{if } b_2 \text{ is desired} \end{cases} \tag{23}$$

The final determination of whether a valid mapping for both α and k exists for the Allais Paradox depends on the combination of preferred belief states chosen by the self-deceptive agent. Only when the preferred belief states are certainty of a_1 and b_2 does there exist a valid space of mappings for α and k as shown in Figure 2.

5. CONCLUSION

Human rationality and decision making are cornerstones in the fields of economics, psychology, and sociology and research into descriptive and psychologically-inspired models of decision making will continue to be a vital factor in any field of research touching on human decision-making capabilities.

In this work, we have presented a formalized framework for modeling the phenomenon of self-deception and wishful thinking within a decision-theoretic framework. Our self-deceptive framework is based on and leverages EU Theory for both the formulation of the desired belief state and the subsequent integration and decision-making process.

While this framework presents a psychologically motivated descriptive framework for self-deception we have also shown that in uncertain environments in which errors of causality and subjective utility evaluations exist, there are situations in which self-deceptive decision making may outperform de-

	b_1	b_2	b_3	b_4
X_7	1	1	0	0
X_8	k	0	k	0

Table 13: Allais Paradox: Outcomes for X_7, X_8

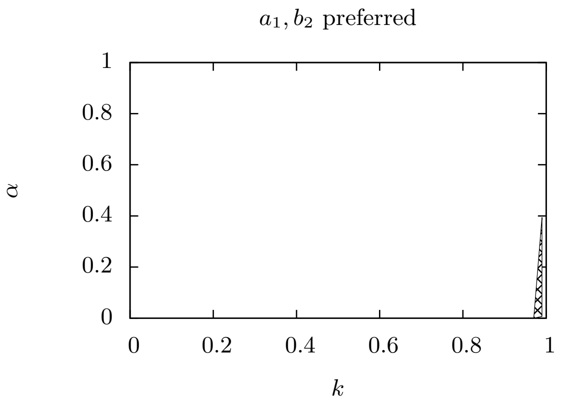

Figure 2: Allais Paradox: Valid α and k Mappings

cisions ascribed by EU maximization. Finally, we discuss the application of our framework with respect to the Certainty Effect and the Allais Paradox, both of which are observed inconsistencies in human decision making with respect to EU Theory and show that our self-deceptive formulation of wishful thinking possesses the descriptive flexibility needed to account for these inconsistencies in a principled fashion.

6. REFERENCES

[1] L. Alloy, L. Abramson, and D. Viscusi. Induced mood and the illusion of control. *Journal of Personality and Social Psychology*, 41(6):1129–1140, 1981.

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

[3] F. Ayeroff and R. Abelson. ESP and ESB: Belief in personal success at mental telepathy. *Journal of Personality and Social Psychology*, 34(2):240–247, 1976.

[4] D. Bell. Regret in decision making under uncertainty. *Operations Research*, 30(5):961–981, 1982.

[5] V. Benassi, P. Sweeney, and G. Drevno. Mind over matter: Perceived success at psychokinesis. *Journal of Personality and Social Psychology*, 37(8):1377, 1979.

[6] 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.

[7] R. Cialdini. *Influence: science and practice.* Allyn and Bacon, 2001.

[8] L. Festinger. *A Theory of Cognitive Dissonance.* Stanford University Press, 1957.

[9] J. Gratch and S. Marsella. A domain-independent framework for modeling emotion. *Cognitive Systems Research*, 5(4):269–306, 2004.

[10] J. Ito, D. Pynadath, and S. Marsella. Modeling Self-deception within a Decision-Theoretic Framework. In *Intelligent Virtual Agents: 8th International Conference, IVA 2008, Tokyo, Japan, September 1-3, 2008, Proceedings*, page 322. Springer-Verlag GmbH, 2008.

[11] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 1979.

[12] Z. Kunda. The case for motivated reasoning. *Psychological Bulletin*, 108(3):480–498, 1990.

[13] Z. Kunda. Motivated Inference: Self-Serving Generation and Evaluation of Causal Theories. *Social Cognition: Key Readings*, 2004.

[14] E. Langer. The illusion of control. 1975.

[15] R. Lazarus. *Emotion and Adaptation.* Oxford University Press, USA, 1991.

[16] G. Loomes and R. Sugden. Regret theory: An alternative theory of rational choice under uncertainty. *Economic Journal*, 92(368):805–824, 1982.

[17] S. Marsella, D. Pynadath, and S. Read. PsychSim: Agent-based modeling of social interactions and influence. In *Proceedings of the International Conference on Cognitive Modeling*, pages 243–248, 2004.

[18] L. Savage. The Foundation of Statistics. *New York*, 1954.

[19] J. von Neumann and O. Morgenstern. Theory of Games and Economic Behavior. *New York*, 1953.