THE INTERPERSONAL EFFECT OF EMOTION IN DECISION-MAKING AND SOCIAL DILEMMAS

by

Celso Miguel de Melo

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(COMPUTER SCIENCE)

Dissertation Committee:
Jonathan Gratch (Chair, Computer Science, USC)
Stacy Marsella (Computer Science, USC)
Milind Tambe (Computer Science, USC)
Peter Carnevale (Marshall School of Business, USC)
Stephen Read (Psychology, USC)
Jeremy Bailenson (Communication, Stanford)

May 2012

Acknowledgments

It is no easy feat having the opportunity to pursue your passion in today's fiercely competitive world. It is even rarer being able to do so surrounded by people willing to offer nothing less than unconditional support. I had the fortune of being in such a position developing this work and would like to acknowledge here some of the people that made that possible. Jonathan Gratch is not an average mentor; it takes a special skill to be able to work with an independent student that prefers to be anywhere, but in the office. His deceivingly 'laid back' style allowed me to develop, unencumbered, the ideas presented here, while having the comfort of worrying only about meeting his highest standards of quality; standards I knew would assure this work a place in the state-of-the-art. It took Peter Carnevale only a glimpse at my virtual agents and I knew, looking at his face, that things would be different from that point onward. With him onboard, and benefiting from his crystal-clear insight and relentless encouragement, this research quickly grew into something much bigger than I had hoped for when I started. A few other scholars deserve a special mention: Milind Tambe, in whose class this whole endeavor began in an attempt to convince him that emotion was missing in the game-theoretic models artificial intelligence was preaching (!); Stacy Marsella, whose insightful feedback in our discussions helped shape this work from start to end; Stephen Read, who helped us very much in sharpening our research methods and establishing a bridge with the social psychology field; Jeremy Bailenson, whose ambition and drive to show the world the importance of virtual reality (and agents) has inspired me; and, last but not least, Ana Paiva, who fought hard to give me the opportunity to develop this work in the United States at USC and, despite being one ocean away, continued to be embarrassingly encouraging. I also thank many other colleagues and friends, both here at the Institute for Creative Technology and back in Portugal, for their admiration and support.

In my personal life, there were a few people that made this journey, not only possible, but fun. I would like to thank, first, my parents for their unconditional love and support in the pursuit of my dreams. Through them, I was also able to maintain my connection to Portugal (my mother) and Mozambique (my father). Constantly being reminded how everybody was proud back home, made me feel like I was doing something that had meaning, not only to myself, but to the family. I could also not have done this without the love of a very special woman. Lital not only rescued me from a social-life free existence in these past years but, also made me feel at home right here, on this side of the Atlantic. There are many others friends and relatives in the United States, Portugal and Mozambique, for whose support I would like to show my deepest gratitude.

Finally, I would like to thank the sponsors that made this work possible: the Fundação para a Ciência e a Tecnologia (FCT) under grant SFRH-BD-39590-2007; the U.S. Army Research, Development, and Engineering Command and the National Science Foundation under grant HS-0713603; the Air Force Office of Scientific Research under grant FA9550-09-1-0507; and, the National Science Foundation under grant IIS-0916858. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

Table of Contents

List of Tables

List of Figures

Abstract

Recent decades have seen increased interest on the role of emotional expression in human-computer interaction. However, despite a growing number of empirical findings reported in the literature, the results are still inconclusive about the mechanism for the social effects of emotion. This dissertation studies how emotion displays in computers impact people's decision-making and proposes a mechanism for such effects based on appraisal theories of emotion. In appraisal theories, emotion displays arise from cognitive appraisal of events with respect to one's goals (e.g., is this event congruent with my goals? Who is responsible for this event?). According to the pattern of appraisals that occurs, different emotions are experienced and displayed. Since displays reflect the agent's intentions through the appraisal process, we argue people infer, from emotion displays, how computer agents are appraising the ongoing interaction and, from this information, make inferences about the agents' intentions. We refer to this theory as reverse appraisal. To support it, several empirical studies are presented where participants engage in the iterated prisoner's dilemma with virtual agents that, though following the same strategy to choose their actions, display emotions in the face that are consistent with either cooperative or competitive goals. The results confirm that emotion displays impact people's decision making and people cooperate more with cooperative agents. Consistent with reverse appraisal's prediction that what is critical for the effects is not the emotion but the underlying appraisals, the results also show that the same display (e.g., a smile) can lead to opposite effects on cooperation depending on the context in which it is shown. A final study shows that people can successfully retrieve, from emotion displays, information about how agents are appraising the ongoing interaction; moreover, these perceptions of appraisal are shown to mediate the effect of emotion displays on perceptions of the agents' likelihood of cooperation. Having established the theoretical

foundations of reverse appraisal, the dissertation presents several computational models of decision-making in the prisoner's dilemma. These models were developed by applying statistical and machine learning techniques on the data collected in the studies. Experimental results show that, as expected, computer models can better replicate human behavior in the original studies if they take into consideration the counterpart's appraisals and emotion displays. In addition, the dissertation also contributes a platform for creating realistic embodied agents that can display emotion. This platform was used in all studies to implement the experimental manipulations and can be used as a research tool to study human-agent and human-human interaction. Finally, the dissertation discusses the implications of reverse appraisal and the reported social effects of emotion for human-computer interaction, artificial intelligence, decision and emotion theory.

Chapter One: Introduction

1.1 Vision

The promise of human-centered computing is to make human-computer interaction as easy, natural and efficient as human-human interaction itself. Seamless integration of computing into the fabric of daily life requires, therefore, computers to be able to understand not only the verbal but, the nonverbal language of humans. Emotional signals are a key component of the latter. The communication and interpretation of emotion displays are skills humans use pervasively to regulate social interaction in personal and professional life. A smile succinctly communicates affiliation; an angry expression conveys disapproval and a threat of retaliation; and, so on. Computer systems of the future need, thus, to be emotionally competent. Aside from recognizing human displays of emotion, these systems need to replicate the *function* of emotion expression. Not necessarily the form. Herbert Simon (1969) points out that, in contrast to the natural sciences which seek to describe intelligence as it is found in nature, the "artificial sciences" seek to describe intelligence as it "ought to be in order to attain goals, and to function." This normative emphasis leads to succinct reconstructions of "messy" biological phenomena so that they may be simulated in a multitude of computer systems that exist today and will exist in the future. The goal is, therefore, to understand and abstract the function of emotion displays. This can only be accomplished by computer systems that reflect deep psychological theories of the social effects of emotion in human life and, do not necessarily replicate surface forms of human expression. Modern computer science, thus, requires a multi-disciplinary perspective. This dissertation adopts such a perspective and contributes to this vision by studying the social effects of emotion in computer systems on one important aspect of people's lives: decision-making.

1.2 Motivation

Emotional skills, especially the ability to recognize and express emotions, are essential for natural communication in humans and, thus, are critical for human-centered computing (Pantic, Pentland, Nijholt, & Huang, 2006; Picard, 1997; Zeng, Pantic, Roisman, & Huang, 2009). A growing number of studies have explored emotion in *embodied agents* (or *virtual humans*) to enhance interaction with computers (Beale & Creed, 2009). Embodied agents are software agents that have virtual bodies and can express themselves through them in the same way people do (Gratch et al., 2002). However, Beale and Creed (2009) emphasize in a recent survey that, despite a large number of empirical studies, it is still unclear how people respond to agents that display emotions and whether they can enhance human-computer interaction. Acknowledging the value of emotion for human-computer interaction is, thus, not sufficient; it is further necessary for computer scientists to understand the psychological theories of emotion in order to effectively design computers that are able to recognize, synthesize and express emotions (Calvo & D'Mello, 2010). This dissertation presents cross-disciplinary work that, building on an appropriate psychological theory, studies how emotional displays in computer agents impact people's decision-making.

The dissertation focuses on one specific type of decision-making situation known as *social dilemmas*. In social dilemmas, people are faced with a decision between pursuing their own self-interest or trusting another person to reach mutual cooperation and maximize joint reward (Kerr, 2011; Kollock, 1998; Komorita & Parks, 1994; Messick, 1983; van Lange, Liebrand, Messick, & Wilke, 1992). In such dilemmas, decision theorists argue that the rational thing for a person to do is act so as to maximize expected utility (Arrow, 1971; Bernoulli, 1738; Friedman & Savage, 1948; Keeney & Raiffa, 1976; von Neumann & Morgenstern, 1944), which corresponds to pursuing the choice that maximizes self-interest. The dilemma is that, if all parties act "rationally", then the collective outcome is worse for all. Researchers were quick to realize that

people do not always follow this narrow view of self-interest and frequently break the assumptions of rational behavior (Allais, 1953; Camerer, 1995; Simon, 1997; Starmer, 2000; Tversky & Kahneman, 1979; Pruitt & Kimmel, 1977). Early research in the behavioral sciences has, in fact, shown many sources of cooperation when people engage in social dilemmas: some people are simply inclined to cooperate (McClintock & Liebrand, 1988); group identity (Kramer & Brewer, 1986); reciprocity (Axelrod, 1984); perception of efficacy of one's contribution (Kerr, 1989); monitoring and sanctioning (Yamagishi, 1986); and, verbal communication (Balliet, 2010; Orbell, van de Kragt, & Dawes, 1988). Recently, researchers started emphasizing the impact of emotion in human decision-making (Blanchette & Richards, 2010; Loewenstein & Lerner, 2003). Contrary to the classical view of emotion as an obstacle to rational decision-making (e.g., Hirschman, 1997; Lefford, 1946), this research emphasizes emotion's potential benefits (Bechara, Damasio, Tranel, & Damasio, 1997; Blanchette, Richards, Melnyk, & Lavda, 2007; Damasio, 1994; Wilson & Schooler, 1991).

Whereas one line of research on the role of emotion in decision-making has emphasized the *intra*personal effect of emotion (Blanchette & Richards, 2010; Loewenstein & Lerner, 2003)–i.e., the impact of one's own emotions in one's decision-making–this dissertation considers emotion's *inter*personal effect–i.e., the impact of another's emotions on one's decision-making. Effectively, several researchers have argued that emotions serve important social functions (Frijda & Mesquita, 1994; Keltner & Haidt, 1999; Keltner & Kring, 1998; Morris & Keltner, 2000; Oatley & Jenkins, 1996). This research emphasizes that emotional expressions are not simple manifestations of internal experience; rather, expressions are other-directed and communicate intentions, desired courses of actions, expectations and behaviors (Bavelas, Black, Lemery, & Mullet, 1986; Fernandez-Dols & Ruiz-Belda, 1995; Keltner & Kring, 1998; Kraut & Johnston, 1979). The expression of emotion has also been theoretically argued to play a significant role in

the emergence of cooperation in social dilemmas (Boone & Buck, 2003; Frank, 1988; Nesse, 1990; Trivers, 1979). Empirically, several studies support the claim that facial displays of emotion influence cooperation (Brown, Palameta, & Moore, 2003; Chapman, Kim, Susskind, & Anderson, 2009; Krumhuber, Manstead, & Kappas, 2007; Scharlemann, Eckel, Kacelnik, & Wilson, 2001; Schug, Matsumoto, Horita, Yamagishi, & Bonnet, 2010).

This dissertation studies the interpersonal effect of emotion when people engage in a social dilemma with embodied agents that display emotion. Previous research has shown that people can treat embodied agents like other people (Nass, Steuer & Tauber, 1994; Reeves & Nass, 1996) and can be socially influenced by them (Blascovich, 2002). Moreover, embodied agents have been argued to be an appropriate research tool for basic human-human interaction research (Blascovich et al., 2002). In our case, participants engaged in a social dilemma with agents that, even though following the same strategy to choose their actions, conveyed different facial emotion displays according to the dilemma outcome. In line with expectations from the behavioral sciences, the results indicated that people's decision to cooperate was, in fact, influenced by emotion displays. For instance, people cooperated more with an agent whose displays reflected mutual cooperation (e.g., smile when both players cooperated) than one whose displays reflected selfishness (e.g., smile when it defected and the participant cooperated).

The focus of the dissertation is, however, on the *mechanism* by which emotion displays influence people's decision-making in a social dilemma. Following the aforementioned social-functions view of emotion, we look at an explanation based on appraisal theories of emotion. In appraisal theories (Ellsworth & Scherer, 2003), emotion displays arise from cognitive appraisal of events with respect to an agent's goals, desires and beliefs (e.g., is this event congruent with my goals? Who is responsible for this event?). According to the pattern of appraisals that occurs, different emotions are experienced and displayed. Since displays reflect the agent's intentions

through the appraisal process, it is also plausible to ask whether people can infer from emotion displays the agent's goals by reversing the appraisal mechanism. We refer to this theory as *reverse appraisal*. The intuition is that if appraisal, abstractly, is a function that maps from <event, mental state> to emotion, reverse appraisal is a function that maps from <event, emotion> to mental state, see Figure 1.1. Empirical evidence is still scarce but in a recent study Hareli and Hess (2010) showed that people could, from expressed emotion, make inferences about the character of the person displaying emotion. So, for instance, a person who reacted with anger to blame was perceived as being more aggressive, self-confident but also as less warm and gentle than a person who reacted with sadness. The empirical studies presented in this dissertation show the plausibility of reverse appraisal as the underlying mechanism for the impact of emotion displays on people's decision-making.

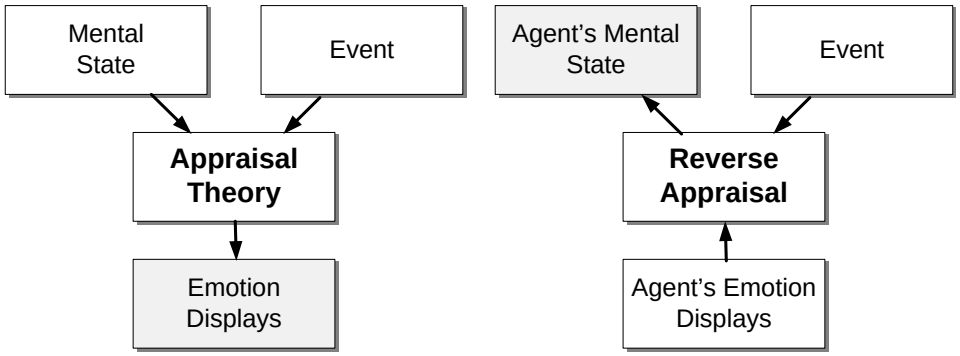


Figure 1.1. Appraisal theory and reverse appraisal.

The dissertation has several implications for the ongoing debate about the impact of emotion in human-computer interaction. Acknowledging that people can treat computers socially (Blascovich, 2002; Blascovich et al., 2002; Gratch et al., 2002; Nass et al., 1994; Reeves & Nass, 1996), researchers predict that emotion expression can have a beneficial impact in human-agent interaction as is seen in human-human interaction. Current research has, thus, focused on showing

that emotion can enhance interaction (Beale & Creed, 2009; Dehn & Van Mulken, 2000), in neglect of understanding the mechanisms by which emotion influences human-agent interaction. It is no surprise, then, that many studies focused on simple comparisons between agents that displayed emotions when compared to agents that did not (Hone, 2006; Klein, Moon, & Picard, 2002; Lim & Aylett, 2007; Lester et al., 1997; Liu & Picard, 2005; Maldonado et al., 2005; Prendinger, Mayer, Mori, & Ishizuka, 2003); and, the few studies that compared agents that expressed different emotions compared simple aspects of emotion and did not frame the results within a broad theory of emotion (Brave, Nass, & Hutchinson, 2005; Gong, 2007). As a result, an incomplete view of the impact of emotion emerges, referred here as the *affective persona effect*, which argues that the mere presence of consistent emotions in agents is sufficient to improve human-computer interaction. This view is reminiscent of the persona effect (Lester et al., 1997; Van Mulken, André, & Muller, 1998), which argues that the mere presence of agents with virtual bodies is sufficient to enhance human-computer interaction. The affective persona effect is, however, at odds with a social-functions view of emotion and appraisal theories, and this dissertation shows that it is not the mere presence of consistent emotion but the *context* under which emotion is expressed and the *information* conveyed by emotion that has the potential to enhance human-computer interaction.

The dissertation also has implications for the study of decision-making in artificial intelligence. In many ways, artificial intelligence has been following the path of the behavioral sciences in the study of decision-making. First, a game-theoretic foundation was established for single-agent decision-making under risk (Russell & Norvig, 2010) and multi-agent decision-making in interactions such as social dilemmas and negotiation (Kraus, 1997; Shoham & Leyton-Brown, 2009; Wooldridge, 2009). Realizing that some of the assumptions in game theory are computationally unreasonable, researchers then turned to, as opposed to optimal rational

solutions, satisfactory solutions (Simon, 1956) and models of bounded rationality (e.g., Aumann, 1997). Then, as interest in human-agent encounters grew, researchers started simulating the ways actual humans decide in human-human encounters (e.g., Lin & Kraus, 2010). However, emotion was largely absent in this endeavor. Despite that emotion had long been argued to be critical to artificial intelligence (Minsky, 1986; Simon, 1967; Sloman & Croucher, 1981), it was only recently that researchers began considering it in their models (Marsella, Gratch, & Petta, 2010). Many systems have, now, attempted to simulate emotion synthesis, the majority of which based on appraisal theories of emotion (Becker-Asano & Wachsmuth, 2008; Dias & Paiva, 2005; Gratch & Marsella, 2004; Wehrle & Scherer, 2001). Some systems have, further, explored the cognitive impact of emotion, in particular, its role in juggling multiple competing goals (Gratch & Marsella, 2004; Scheutz & Schermerhorn, 2009; Scheutz & Sloman, 2001; Staller & Petta, 2001). However, these systems tend to focus on the intrapersonal effect of emotion in decision-making and, neglect the interpersonal impact of emotion in decision-making. The dissertation advances the study of decision-making in artificial intelligence by, first, exploring the theory of interpersonal effect of emotion in decision-making and, second, proposing computer models of decision-making in social dilemmas that are influenced by the counterpart's emotion displays. Moreover, the reverse appraisal proposal emphasizes that what matters for the social effects of emotion are not the emotions per se, but the information conveyed by emotion. This is the abstract function of emotion displays and, thus, the key to extend our results to computational systems that go far beyond virtual agents and, besides enhancing human-computer interaction, improve agent-agent interaction.

Finally, the dissertation has implications for decision and emotion theory. The empirical studies described here report social effects of discrete emotions on people's decision-making in a social dilemma. This evidence emphasizes that emotions serve important social functions such as

communicating one's intentions to others in social decision-making (Frijda & Mesquita, 1994; Keltner & Kring, 1998; Morris & Keltner, 2000; Van Kleef, De Dreu, & Manstead, 2010). Reverse appraisal proposes further a specific mechanism for these functions whereby emotion displays convey information about how the counterpart is appraising the ongoing interaction which, then, lead to inferences about the counterpart's mental state, in particular his or her likelihood of cooperation in a social dilemma. The dissertation also makes a methodological contribution for the study of the social effects of emotion on decision-making. In line with the idea that virtual agents can be used for basic social psychology research (Blascovich et al., 2002), all studies used agents to create the experimental manipulations. Using virtual agents allowed precise experimental control, low-cost, easy, replicable, and incremental research. Moreover, aside from extending current knowledge, the findings were compatible with (and, in some cases, replicated) previous findings from the behavioral sciences regarding human-human interaction, suggesting people interact naturally with these agents.

1.3 Objectives

The dissertation's goals were to:

- Show that people's decision making is influenced by computer agents' emotion displays in social dilemmas;
- Propose the reverse appraisal mechanism for the interpersonal effect of emotion in people's decision-making in social dilemmas;
- Develop computational models of decision-making that accounted for the interpersonal effect of emotion in social dilemmas.

1.4 Approach and Hypotheses

To show the effect of emotion displays in decision-making, a first empirical study is presented where participants played the iterated prisoner's dilemma for several rounds with embodied agents. In this study, the agents followed the same strategy to choose their actions–tit-for-tat (Axelrod, 1984)–but conveyed facial emotion displays consistent with different social value orientations (e.g., for the outcome where the participant was exploited by the agent, a cooperative agent showed guilt, whereas a competitive agent smiled). The main dependent measure was cooperation rate over all rounds. The hypothesis was that different patterns of facial display would influence people's decision-making and, thus, lead to effects on cooperation rate.

A second study continued unpacking the mechanism for the interpersonal effect of emotion in the prisoner's dilemma. Appraisal theories argue that the cause of emotion is context-dependent (Ellsworth & Scherer, 2003) and, thus, the same display can occur in rather distinct situations. Building on this insight, the new study compared agents which, as in the first study, followed the same strategy but, unlike the first study, differed only in the context in which emotions were displayed. For instance, a cooperative agent showed a smile in mutual cooperation, whereas a competitive agent showed a smile when the agent exploited the participant. The hypothesis was that, as predicted by appraisal theories, people would interpret differently the same smile and cooperate differently with the agents.

A third set of experiments was, then, conducted to provide evidence for reverse appraisal. Following findings of Hareli and Hess (2010), people were hypothesized to use emotional displays to infer beliefs, desires and intentions of their social partners essentially by reversing the appraisal mechanism. To see if this mechanism explained the prior prisoner's dilemma findings, participants were asked to imagine playing the dilemma with different agents; participants were always told the same outcome occurred but were shown videos of different emotional reactions

from the agent and were, then, queried about how they thought the agent was appraising the situation and how likely it was to cooperate in the future. A statistical analysis of mediation (Baron & Kenny, 1986; Preacher & Hayes, 2008) was conducted to test whether appraisal variables (e.g., conduciveness to goals and blameworthiness) mediated the effect of emotion displays on people's perception of how likely the agent was to cooperate in the future. The hypothesis was that appraisal variables would mediate the interpersonal effect of emotion. To further test the mediating role of appraisals, a follow-up experiment explicitly manipulated appraisals and measured the effect on people's perception of how cooperative the agent was. The manipulation consisted of having the agents, instead of showing facial displays of emotion, express how they were appraising the outcome through text (e.g., "I really don't like this outcome and I blame you for it"). The hypothesis was that, in line with the reverse appraisal proposal, expression of appraisals would lead to effects on perception of the agent's cooperativeness that were consistent with findings in the previous experiment.

Having established the theoretical foundations for reverse appraisal, we began exploring computational models for decision-making in the prisoner's dilemma. The first model was developed using a statistical technique–maximum likelihood estimation (Alpaydin, 2010)–and the data from the first two studies. Several variants of this model were explored including one that predicted cooperation given information about the outcome alone and another given information about the outcome and the emotion displayed by the counterpart. The hypothesis was that the model that considered emotion would replicate better the empirical data in the original studies than the model that did not. The second model sought to show the value of integrating appraisals into computational models of decision-making. To accomplish this, Bayesian learning (Alpaydin, 2010) was used to learn the model from data in the last study. Again, several variants were explored including one that considered outcome and emotion displays and another that considered

outcome, emotion displays and appraisals. Models were evaluated with respect to their ability to replicate the empirical data in the original studies. Because appraisal theories advocate a shared appraisal structure for emotions, models that consider appraisals can learn things about unseen emotions as long as training data is provided for emotions that share appraisals with the unseen emotions. Thus, we hypothesized that models that considered appraisals would have better accuracy than models that did not over test sets which included emotions not seen in the training set. Lastly, there are situations where people express how they are appraising a situation without resorting to emotion expression. An obvious example is when people convey verbally their attitudes toward an event. The data collected in the experiment where people conveyed appraisals through text is a case in point. This dataset could, thus, be used to test our final hypothesis: Models that considered appraisals were accurate even when no emotion was shown.

1.5 Contributions

The dissertation contributions are:

- In line with predictions from the behavioral sciences regarding the role of nonverbal cues on people's decision-making, empirical evidence that people's decision to cooperate in social dilemmas is influenced by computer agents that display emotion;
- In line with a social-functions view of emotion, evidence that people infer the agent's intentions, desires and beliefs from its emotion displays. In particular, evidence that these inferences are accomplished through reverse appraisal, i.e., people infer from emotion displays how the agent is appraising the ongoing interaction which, in turn, leads to inferences about the agent's likelihood of cooperation;
- Computer models for decision-making in a social dilemma that take into account the counterpart's emotion displays;

- A novel paradigm for the investigation of human-human and human-agent interaction based on experimental games with emotionally expressive virtual humans and a virtual humans platform that supports this framework.

1.6 Dissertation Structure

The rest of the dissertation is structured as follows: Chapter 2 overviews the relevant literature in decision-making, emotion theory, human-computer interaction and artificial intelligence; Chapter 3 describes in detail the empirical studies that were conducted to support the reverse appraisal theory; Chapter 4 presents the computer models of decision-making; and, Chapter 5 discusses the results, overviews the contributions and implications, and discusses future work. The Appendix describes the virtual humans research tool that supports our research paradigm and is used in the empirical studies.

Chapter Two: Background

This chapter begins by overviewing theory in decision-making and emotion that is relevant to understand the dissertation's reverse appraisal proposal and place it within existent research in the behavioral sciences. Section 2.1 describes early mathematical frameworks that were used to study optimal, or "rational", decision-making. Section 2.2 proceeds to describe research that showed that people consistently deviate from the rational predictions in these mathematical frameworks. The section also introduces emotion as one of the causes for such deviations. Sections 2.3 and 2.4 then detail research on the impact of emotion in decision-making. Whereas the former summarizes the extensive body of research that focused on the impact of experienced emotion on one's own decision-making (intrapersonal), the latter reports relatively more recent research on the impact of expressed emotion on others' decision-making (interpersonal). Finally, Section 2.5 reviews appraisal theories and introduces the reverse appraisal proposal. Having reviewed the theoretical foundations, Sections 2.6 and 2.7 focus on computational systems and describe, respectively, related work on the impact of emotion displays on decision-making in human-computer interaction and artificial intelligence. These sections also clarify the novel contribution of reverse appraisal in comparison to existent work in computer science.

2.1 Rational Decision-Making

Research in rational decision-making investigates how people estimate the likelihood of different outcomes and choose between different options (Doyle, 1998). Game theory is a branch of mathematics devoted to understanding rational decision-making (Osborne & Rubinstein, 1994). A central assumption in game theory is that decision-makers are "rational". Rationality, in turn, hinges on the principle of *maximum expected utility* which says that a rational decision-maker

should choose the action that maximizes its expected utility. The expected utility of an action is the average utility of all the outcomes that might occur, weighted by the probability of each outcome. The principle of maximum expected utility is argued to lead to rational behavior because it leads the decision-maker to act consistently with its preferences, just as long as certain constraints over the preferences–the axioms of utility theory (ordering, continuity and independence)–are obeyed (von Neumann & Morgenstern, 1944).

The models of game theory are abstract representations of real-life situations where decision-makers interact strategically (Osborne & Rubinstein, 1994). Game theoretic models can be subdivided into strategic games, extensive games (with or without perfect information) and coalition games. In strategic and extensive games the primitives are actions of individual players and in coalition games the primitives are actions of groups of players. In strategic games, players choose their actions once and for all, and these choices are made simultaneously. These games are represented by a matrix that shows the players, strategies and payoffs. Extensive games, in contrast, formalize a time sequence of actions by each player. These games are represented by a tree where each node represents a point of choice for a player. A general model of an extensive game allows each player, when making its choices, to be perfectly or imperfectly informed about what has happened in the past. The dissertation focuses on one strategic game–the prisoner's dilemma–and, its repeated version–the iterated prisoner's dilemma–which can be seen as an imperfect-information extensive game.

The prisoner's dilemma (Poundstone, 1993) is classically described as follows:

> Two suspects are arrested by the police. The police have insufficient evidence for a conviction, and, having separated the prisoners, visit each of them to offer the same deal. If one testifies for the prosecution against the other (defects) and the other remains silent (cooperates), the defector goes free and the silent accomplice receives the full 10-year

> sentence. If both remain silent, both prisoners are sentenced to only 6 months in jail for a minor charge. If each betrays the other, each receives a 5-year sentence. Each prisoner must choose to betray the other or to remain silent. Each one is assured that the other would not know about the betrayal before the end of the investigation.

The canonical payoff matrix for this game is shown in Table 2.1, where *T* stands for *Temptation to defect*, *R* for *Reward for mutual cooperation*, *P* for *Punishment for mutual defection* and *S* for *Sucker's payoff*. To be defined as prisoner's dilemma, the following inequalities must hold: T > R > P > S. Applying the canonical form to the story described above we get: T = goes free; R = 6 months; P = 5 years; S = 10 years. In a standard game-theoretic analysis of this game, the rational thing for each player to do is defect. Effectively, if prisoner A believes prisoner B is going to defect, then the best it can do is defect too; if prisoner A believes prisoner B is going to cooperate, then the best it can do is again defect. By symmetry, prisoner B should also decide to defect. Mutual defection is thus a *Nash equilibrium* for this game, i.e., no player can gain by unilaterally moving away from the equilibrium choices. However, mutual defection is a *Pareto*-suboptimal solution, i.e., there is another solution to the game–mutual cooperation–which both players would prefer. This is why the prisoner's dilemma is referred to as a social dilemma.

Table 2.1. Canonical payoff matrix for the prisoner's dilemma.

		Prisoner B	
		Cooperates	Defects
Prisoner A	Cooperates	Prisoner A: R Prisoner B: R	Prisoner A: S Prisoner B: T
	Defects	Prisoner A: T Prisoner B: S	Prisoner A: P Prisoner B: P

Social dilemmas are situations in which individual rationality leads to collective irrationality (Kerr, 2011; Kollock, 1998). In the prisoner's dilemma, if every player follows the individually

rational strategy, then mutual defection should result. However, this outcome is collectively worse than mutual cooperation. In the iterated version of the prisoner's dilemma, players play several rounds of the game but, each player is informed about the action the other player chose in the previous round. In this case, it is also common to impose an extra restriction on the canonical payoff matrix: $2R > S + T$. The game-theoretical solution to the iterated prisoner's dilemma depends on whether the number of rounds is finite or infinite. If the number of rounds is infinite, then the rational solution becomes mutual cooperation. The aforementioned extra restriction assures that mutual cooperation is better than alternating between player A exploiting player B (i.e., player A defecting and player B cooperating) and player B exploiting player A. However, if the number of rounds is finite, then the rational solution is mutual defection in every round. The reasoning is that the last round is effectively a one-shot prisoner's dilemma game and, thus, players should defect. But then, by induction, players should defect in every round. In this dissertation we explore the iterated prisoner's dilemma with a finite number of rounds.

2.2 Human Decision-Making

Researchers quickly realized that people frequently violate game-theoretic predictions of rational behavior (for detailed reviews see: Camerer, 1995; Starmer, 2000). First evidence contradicting the principle of maximum expected utility came from Maurice Allais (1953). In his experiment, participants are asked to choose between two options: the first option gives 1 million Francs guaranteed; the second gives 5 million with a probability of 0.1, 1 million with a probability of 0.89 and, otherwise, nothing. People consistently preferred the first option over the second, despite that the second has a higher expected utility. Herbert Simon (1956) argues that deviations from rationality happen because humans lack the cognitive resources to calculate optimal rational solutions and, instead, seek satisfactory solutions. This led researchers to explore several models of bounded rationality. Bounded rationality (Simon, 1997) studies decision-making under the

assumption that people are limited by the information they possess, their cognitive abilities and time. Tversky and Kahneman's famous prospect theory (1979, 1981) also describes several cognitive heuristics people use when making decisions.

Systematic deviations of rational behavior have also been found in social dilemmas (Kollock, 1998; Pruitt & Kimmel, 1977). In the (finite) iterated prisoner's dilemma, contrary to the game-theoretic prediction of mutual defection in every round, early research in the behavioral sciences has shown that people actually achieve mutual cooperation often and, several sources of cooperation have been identified (a) some people are simply inclined to cooperate (McClintock & Liebrand, 1988), (b) group identity (Kollock, 1998b; Kramer & Brewer, 1986), (c) reciprocity (Axelrod, 1984), (d) perception of efficacy of one's contribution (Kerr, 1989), (e) monitoring and sanctioning (Winett, Kagel, Battalio, & Winkler, 1978; Yamagishi, 1986) and, (f) verbal communication (Balliet, 2010; Jerdee & Rosen, 1974; Orbell et al., 1988).

The research reviewed in this section provides strong evidence that people's decision-making does *not* strictly follow game-theoretic predictions of rational behavior. Though bounded rationality, cognitive heuristics and the aforementioned sources of cooperation explain some of the variance in people's decision-making with respect to rational behavior, there is one more factor that accounts for some of this variance and that has been the subject of much recent attention and is also the focus of this dissertation: emotion. The next two sections address the impact of emotion in decision-making: subsection 2.3 reviews research on the intrapersonal effect of emotion; and, subsection 2.4 reviews research on the interpersonal effect of emotion.

2.3 Intrapersonal Effect of Emotion on Decision-Making

Reviewing earlier views of emotion, Loewenstein and Lerner (2003) report that "Throughout most of recorded human intellectual history, emotions have been viewed largely in negative terms" and that "philosophers focused mainly on the role played by emotions in self-control

problems–on the propensity for emotions to override reason, deliberation, or self-interest" (p. 633). Effectively, there are consistent reports and evidence that people experiencing high-intensity emotions can lose control and act against their own long-term self-interest (Baumeister, Heatherton, & Tice, 1994; Bazerman, Tenbrunsel, & Wade-Benzoni, 1998; Loewenstein, 1996). Research in the reasoning literature also provides evidence that when affect is induced on people (Melton, 1995; Oaksford, Morris, Grainger, & Williams, 1996; Palfai & Salovey, 1993) or when reasoning over affective contents (Blanchette & Richards, 2004; Lefford, 1946), people's ability to make valid logical inferences is negatively affected.

However, recent results emphasize the positive influence of emotion in decision-making. Empirical evidence from neuroscience comes from Damasio and colleagues (Bechara, Damasio, Damasio, & Anderson, 1994; Bechara et al., 1997; Damasio, 1994). Their work shows that patients with damage to the prefrontal cortex, causing minimal cognitive but major emotional damage, have difficulty making decisions that come easily to healthy adults. In another line of research, empirical evidence from Wilson and colleagues (Nisbett & Wilson, 1977; Wilson & Schooler, 1991; Wilson & Brekke, 1994) suggests that when participants are asked to justify their preferences over certain objects, they often end up making a choice which does not reflect their true preference. The argument is that justifying their choice leads to a "cognitive" decision that neglects the input of emotion and, thus, ignores "gut feelings" that reflect the participants' true disposition toward the object. Finally, recent studies in the reasoning literature have shown that, in contrast to earlier work in the field, people reason more logically about personal emotional experiences than about neutral contents (Blanchette & Campbell, 2005; Blanchette, Richards, & Cross, 2007; Blanchette, Richards, Melnyk et al., 2007).

Overall, there is now considerable research on the intrapersonal effect of emotion in decision making, i.e., the impact of felt emotion on one's own decision-making (for detailed reviews see:

Blanchette & Richards, 2010; Loewenstein & Lerner, 2003). However, recently, there has been a growing interest on the interpersonal effect of emotion in decision-making, i.e., the impact of other's emotions on one's own decision-making.

2.4 Interpersonal Effect of Emotion on Decision-Making

Several researchers have argued that emotions serve important social functions (Frijda & Mesquita, 1994; Keltner & Haidt, 1999; Keltner & Kring, 1998; Morris & Keltner, 2000; Oatley & Jenkins, 1996). This view emerges from studies documenting how interpersonal problems elicit emotions (e.g., Averill, 1980; Keltner & Buswell, 1997; Miller & Leary, 1992) and how those expressions trigger interpersonal interactions that address the originating problem (Hazan & Shaver, 1987; Johnson-Laird & Oatley, 1992; Lutz & White, 1986; Nesse, 1990). These studies emphasize that emotional expressions are not simple manifestations of internal experience; rather, expressions are other-directed and communicate intentions, desired courses of actions, expectations and behaviors (Bavelas et al., 1986; Fernandez-Dols & Ruiz-Belda, 1995; Keltner & Kring, 1998; Kraut & Johnston, 1979). Regarding means of expression, a common channel used to communicate emotion to others is facial expression (Keltner & Ekman, 2000).

The expression of emotion has also been argued to play a significant role in the emergence of cooperation (Boone & Buck, 2003; Frank, 1988; Nesse, 1990; Trivers, 1971; Van Kleef et al., 2010). Trivers (1971) highlights the role of anger and gratitude on the evolution of cooperative alliances, arguing emotions are a critical element of stable long-term relationships. Frank (1988, 2004) argues that commitment is critical for the success of any business relationship and, that emotions convey this commitment to others and motivate the self to transcend self-interest in order to preserve the relationship and promote cooperation. Boone and Buck (2003) propose an evolutionary argument that emotional expressivity–positive or negative–is associated with cooperative individuals. The idea is that emotionally expressive people are likely to reveal (or

leak) their true intentions and, thus, unlikely to attempt to cheat. Complementary, a defector trying to impersonate a cooperator would have a hard time trying to produce correctly all patterns for positive and negative emotion displays. Nesse (1990) speculates about how emotions help individuals cooperate when engaged in a situation similar to the prisoner's dilemma. For each cell in the prisoner's dilemma payoff matrix, he proposes that certain emotions should be communicated (e.g., in mutual cooperation, both players should express friendship, love, obligation and pride). The studies described in the dissertation relate to Nesse's proposal in that the experimental manipulation consists of changing the pattern of emotions expressed in each cell, and testing how different patterns influence people's decision-making; however, rather than following an evolutionary argument, the assignment of emotions to each cell is based on concrete predictions from appraisal theories (see Section 2.5). Finally, building on empirical results on negotiation, Van Kleef et al. (2010) structure the social effects of emotion on cooperation according to affective and inferential processes. Affective processes occur when emotion displays elicit affective reactions in others (e.g., emotional contagion; Hatfield, Cacioppo, & Rapson, 1994). Inferential processes occur when people interpret emotion displays as information. The argument is that, since specific emotions arise in specific situations, emotion displays provide differentiated information about how the other person regards the situation. Despite not providing explicit predictions for the prisoner's dilemma, their work relates to this dissertation in that this work also focuses on the information people retrieve from emotion displays. However, the dissertation goes further in proposing that this information pertains to how the counterpart is appraising the situation. Van Kleef et al. also propose that the 'perceived competitiveness of the situation' moderates the social effect of emotion. The idea is that the same emotion can have different effects according to how competitive or cooperative a situation is. This dissertation also emphasizes that context is important; however, rather than focusing on how inherently

competitive a situation is (external), the dissertation emphasizes that what determines how competitive a situation is depends on how the individual appraises the ongoing interaction (internal).

Aside from the theoretical accounts described in the previous paragraph, there has been empirical research exploring the impact of facial displays of emotion on emergence of cooperation. Many studies have shown that cooperative individuals display higher levels of positive emotion than non-cooperators (Brown et al., 2003; Mehu, Grammer, & Dunbar, 2007; Scharlemann et al., 2001); de Jong, Peters, de Cremer and Vranken (2002) provide evidence that when the other party defected in the prisoner's dilemma and blushes, trustworthiness–argued to be an important precursor in the development of cooperation (Ross & LaCroix, 1996)–decreased if the motive for blushing was ambiguous; Krumhuber et al. (2007) showed that the dynamics of facial displays were relevant for the perception of trustworthiness; Chapman et al. (2009) presented evidence that disgust could reveal pro-social tendencies in certain situations; finally, Schug et al. (2010) showed evidence that supported Boone and Buck's (2003) theory that cooperators express more emotion–positive or negative–than non-cooperators. Some studies have focused, instead, on verbal communication of emotion in social dilemmas. Wubben, De Cremer, and van Dijk (2009) showed that, in a 2-person dilemma, communication of disappointment was more likely to promote cooperation than communication of anger. Wubben, De Cremer, and van Dijk (2008) also showed that, in a public goods dilemma, communication of anger, especially by a wealthy member, was more likely to lead the targeted member to quit the group than communication of guilt.

2.5 Appraisal Theories and Reverse Appraisal

Despite the growing body of empirical results on the social effects of emotion in decision-making, the mechanism for these effects is still not well understood. According to the

aforementioned social-functions view, emotion displays communicate information about the individual's intentions. However, what is the information conveyed in emotion displays and how do people retrieve it? We look for an answer to these questions in appraisal theories of emotion. In appraisal theories (Ellsworth & Scherer, 2003), emotion arises from cognitive appraisal of events with respect to the person's goals, desires and beliefs (e.g., is this event congruent with my goals? Who is responsible for this event?). According to the pattern of appraisals that occurs, different emotions are experienced. Thus, a specific event (e.g., the sudden appearance of a bear) does not define by itself which emotion an individual should experience. It is the appraisal of the situation with respect to the individual's goals that defines which emotion will be experienced (e.g., a picnic-goer might experience fear at the sight of a bear but a hunter might experience joy).

Though several appraisal theories have been proposed (Frijda, 1986; Ortony, Clore, & Collins, 1988; Roseman, 2001; Lazarus, 1991; Scherer, 2001), there tends to be agreement on the underlying appraisal dimensions (Ellsworth & Scherer, 2003). The most basic dimensions are perception of *novelty* (with respect to the level of habituation) and the *intrinsic pleasantness* or valence of the stimulus. These appraisals tend to occur in a highly automatic fashion. The next appraisal dimension relates to *goal significance*, i.e., whether the event is relevant to the individual's goals or not. Goal significance is usually subdivided into three appraisals: *conduciveness*, *certainty* and *urgency*. Conduciveness refers to whether the event is consistent with the individual's goals. Certainty refers to the probability of the event actually occurring. Regarding urgency, the more important the goals, the more urgent immediate action becomes. A third appraisal variable is *agency*, i.e., who is responsible for the event. A fourth appraisal variable refers to *coping potential*, i.e., the evaluation of one's ability to deal with the situation. Coping potential is, in turn, subdivided into three appraisals: *control*, *power* and *adjustment*. Control refers to how well an event or its outcomes can be influenced or controlled (by the self,

others or nature). If the situation is controllable, the outcome depends on one's own power to exert control or to recruit others to help. Adjustment concerns the individual's capacity to adapt to changing conditions in the environment, which is particularly important if the individual has no power over the situation. Finally, the last appraisal, *norm compatibility*, recognizes that people live in a social context and assesses how much the event conforms to society's norms. Table 2.2 summarizes how these appraisals relate to joy, anger, sadness and guilt–the emotions explored in the dissertation's empirical studies. In general, predictions tend to be consistent across theories: joy occurs when the event is conducive to one's goals; anger occurs when the event is not conducive to one's goals, is caused by another agent and one has power/control over it; sadness occurs when the event is not conducive to one's goals; guilt occurs when the event is not conducive to one's goals, is caused by the self and is not norm compatible.

Table 2.2. Appraisal patterns postulated by different appraisal theories.

Appraisal	Joy				Anger				Sadness				Guilt[a]		
	Sc	Rs	OCC	ES	Sc	Rs	OCC	ES	Sc	Rs	OCC	ES	Sc	Rs	OCC
Novelty	low	-	-	high	high	-	-	high	low	-	-	low	-	-	-
Pleasantness	high	high	high	high	-	low	low	-	low	low	low	low	-	low	low
Goal significance															
Conduciveness	yes	yes	yes	yes	no	no	no	no	no	no	no	no	-	no	no
Certainty	high	high	high	high	high	-	high	high	high	high	high	high	high	-	high
Urgency	low	-	-	low	high	-	-	high	low	-	-	low	high	-	-
Agency	-	-	-	-	other	other	other	other	-	-	-	-	self	self	self
Coping Potential															
Control	-	low	-	high	high	high	-	high	low	low	-	low	-	high	-
Power	-	low	-	high	high	high	-	high	low	low	-	low	-	high	-
Adjustment	med	-	-	high	high	-	-	high	med	-	-	med	med	-	-
Norm Compatibility	-	-	-	high	low	-	-	low	-	-	-	-	low	-	-

Note. Sc = Scherer (2001); Rs = Roseman (2001); OCC = Ortony et al. (1990); ES = Ellsworth & Scherer (2003).). Entries filled with a dash (-) mean that the theory makes no prediction for the appraisal variable in the case of that emotion or, that the theory does not consider that appraisal variable.

[a] Ellsworth and Scherer (2003) do not present an explicit prediction for guilt.

Since emotion displays reflect the agent's goals through the appraisal process, we argue it is plausible for people to infer from emotion displays the agent's intentions by reversing the

appraisal mechanism. According to this proposal, what people retrieve from emotion displays pertains to information about how the counterpart is appraising the ongoing interaction; this information about appraisals, then, supports further inferences about the counterpart's intentions. Empirical evidence for this proposal is still scarce but, in a recent study, Scherer and Grandjean (2008) show that people are able to make appropriate inferences about appraisals from photos of facial expression of emotion. In another study, Hareli and Hess (2010) show that people can, from expressed emotion, make inferences about the character of the person displaying emotion. So, for instance, a person who reacted with anger to blame was perceived as being more aggressive, self-confident but also as less warm and gentle than a person who reacted with sadness. Moreover, Hareli and Hess show that these perceptions are mediated by perceived appraisals. In a related line of research, Manstead and Fischer (2001) introduce social appraisal theory to emphasize how the appraisals of others can impact one's own emotion. For instance, two people watching a funny movie together will smile more than when watching the movie alone. Manstead and Fischer acknowledge that other people can, of course, be the object of "regular" appraisals (e.g., regarding blameworthiness) but the point is that specific "social" appraisals occur because people care about how others react to situations. Though social appraisal also presupposes people can make inferences about others' appraisals from their emotion displays, the focus of social appraisal is on how these inferences influence the self's experience of emotion; in contrast, reverse appraisal focuses on how inferences about others' appraisals lead to inferences about the other's state of mind. Complementing this preliminary evidence, the dissertation presents several studies that support the reverse appraisal proposal; in particular, the dissertation presents evidence that, when engaged in a social dilemma, people are capable of inferring, from emotion displays, how the counterpart is appraising the ongoing interaction and, from this information, further infer what the counterpart's intentions are.

2.6 Emotion, Decision-Making and Human-Computer Interaction

With the goal of creating more natural and effective human-agent interaction, artificial intelligence and human-computer interaction researchers started exploring embodied agents (or virtual humans) (Gratch et al., 2002). Embodied agents are agents that have virtual bodies and can express themselves through them in the same way people do. Acknowledging that people can treat embodied agents like other people (Nass et al., 1994; Reeves & Nass, 1996) and that people can be influenced by them (Blascovich, 2002; Blascovich et al., 2002), researchers attempted to create agents that display emotions in ways that are consistent with displays people show in daily life. However, the current focus of research has been on showing that emotion can enhance interaction (Beale & Creed, 2009; Dehn & Van Mulken, 2000), rather than in understanding the mechanisms by which emotion influences human-agent interaction. Thus, many studies focused on simple comparisons between agents that displayed emotions when compared to agents that did not (Hone, 2006; Liu & Picard, 2005; Klein et al., 2002; Lester et al., 1997; Lim & Aylett, 2007; Maldonado et al., 2005; Prendinger et al., 2003); some studies compared agents that displayed consistent versus inconsistent emotions (Berry, Butler, & De Rosis, 2005; Creed & Beale, 2008), and naturally concluded that people prefer agents that display consistent emotions; and, the few studies that compared agents that express different emotions compared simple aspects of emotion and did not frame the results within a broad theory of emotion: Gong (2007) showed that people preferred an agent that displayed positive emotions to one that displayed negative emotions, independently of context; and, Brave et al. (2005) showed that people preferred agents that displayed other-empathetic emotions to agents that displayed self-empathetic emotions. As a result, an incomplete view of the impact of emotion in embodied agents emerges, which we refer to as the affective persona effect, that argues that the mere presence of consistent emotions in embodied agents is sufficient to improve human-computer interaction. This view can be seen as a

straightforward extension of the persona effect (Lester et al., 1997; Van Mulken et al., 1998), which argues that the mere presence of embodied agents is sufficient to enhance human-computer interaction, to the case of agents that display emotions.

In line with the affective persona effect, we have shown in the past that people cooperate more with an embodied agent that displayed emotions than one that did not (de Melo, Zheng, & Gratch, 2009). In this study, participants played the iterated prisoner's dilemma with two agents that followed the same strategy to choose their actions (tit-for-tat) but, one displayed emotions consistent with a goal of mutual cooperation (e.g., joy when both players cooperate) whereas the other showed no emotion. The results revealed that participants cooperated significantly more with the cooperative than the control agent. Though compatible with the affective persona proposition, we were not satisfied with the argument that the mere presence of emotion in the cooperative agent was sufficient to explain the results. We believe that the information conveyed by the specific emotions, in the appropriate context, was crucial for the results. This dissertation presents a series of studies that complement these preliminary findings and show the insufficiency of the affective persona effect to explain the effect of emotion displays on cooperation in a social dilemma. Moreover, the dissertation proposes that reverse appraisal, rather than the simple presence of emotion, explains how emotion expression can potentially enhance human-machine interaction.

2.7 Emotion, Decision-Making and Artificial Intelligence

Shoham and Leyton-Brown (2008) write in their multiagent systems textbook "While the area of multiagent systems is not synonymous with game theory, there is no question that game theory is a key tool to master within the field" (p. xiv). Russell and Norvig (2010) in their artificial intelligence textbook write "(…) the MEU [maximum expected utility] principle could be seen as defining all of AI" (p. 611). These two passages emphasize the central role game theory currently

assumes in artificial intelligence and, in particular, in the multiagent systems field. Effectively, two major thrusts of research in these fields have been (1) applying game-theoretic models to situations where self-interested agents have to interact with each other, and (2) finding tractable algorithmic implementations of game-theoretic solution concepts.

However, as emphasized in Section 2.2, game theory models tend to be normative rather than descriptive, i.e., they describe how people *should* act instead of how they actually do. Thus, game-theoretic models seem more appropriate for agent-agent interaction than human-agent interaction. Nevertheless, there has been work in artificial intelligence that attempted to simulate human decision-making such as the work on finite-state automata that simulates bounded rationality (Aumann, 1997) and the work on human-agent automated negotiation (Lin & Kraus, 2010). However, though being closer to replicating how humans decide, this work still neglects the pervasive role of emotion in decision-making.

The view that emotion is critical to artificial intelligence is not new (Minsky, 1986; Simon, 1967; Sloman & Croucher, 1981) but it was only recently that researchers began incorporating emotion into their models (Marsella et al., 2010). Many systems have, now, attempted to simulate emotion synthesis, the majority of which based on appraisal theories of emotion (Becker-Asano & Wachsmuth, 2008; Dias & Paiva, 2005; Gratch & Marsella, 2004; Wehrle & Scherer, 2001). Some systems have, further, explored the cognitive impact of emotion, in particular, its role in juggling multiple competing goals (Gratch & Marsella, 2004; Scheutz & Schermerhorn, 2009; Scheutz & Sloman, 2001; Staller & Petta, 2001). However, these systems tend to focus on the intrapersonal effect of emotion in decision-making and, neglect the interpersonal impact of emotion in decision-making. The dissertation addresses this limitation in the literature and proposes the reverse appraisal theory for the interpersonal effect of emotion; moreover,

computational models for decision-making in the prisoner's dilemma are presented that take into account the information provided by the counterpart's emotion displays.

Chapter Three: Theory

This chapter presents empirical evidence for the reverse appraisal proposal. We first show that people are able to identify, from appropriate emotion displays, how likely are computer agents to cooperate in the future. To accomplish this we designed agents which emotion displays reflected appraisals compatible with specific social value orientations: cooperative or competitive. For instance, when the agent exploited the participant, an *expressively cooperative* agent would show guilt because it appraised the outcome as obstructive to its goals and blamed itself for it; in contrast, for the same outcome, an *expressively competitive* agent would smile because it appraised the outcome as very positive. In a first study, people engaged with the cooperative[1] and competitive agents in the iterated prisoner's dilemma. The results confirmed our expectation that people are able to identify a cooperator from emotion displays and, therefore, cooperate more with it than with a non-cooperator. To emphasize that what is critical for these behavioral effects is not the emotion display per se but the underlying appraisals, we then compared two new agents that displayed the same emotions but, in different contexts. Thus, in a second study, we compared a cooperative agent that smiled in mutual cooperation with a competitive agent that smiled when it exploited the participant; otherwise, the agents were the same. In line with expectations from appraisal theories, participants cooperated more with the cooperative than the competitive agent. We then proceeded to show that people retrieve, from emotion displays, information about how the agents are appraising the ongoing interaction; this information, in turn, leads to inferences about the agents' propensity for cooperation. To accomplish this, in a third study, we had participants imagine playing the prisoner's dilemma with emotional agents. They were told certain outcomes occurred and shown videos of how the agent reacted emotionally; then,

[1] Whenever the context is clear, agents are referred to without the "expressively" adverb.

participants were asked about how the agent was appraising the interaction and, how likely it was to cooperate in the future. The results confirmed that people were able to retrieve information about the agents' appraisals from emotion displays and, subsequently, make inferences about the agents' likelihood of cooperation. Moreover, the results showed that perceptions of appraisal mediated the effect of emotion displays on perception of the agents' cooperativeness. Finally, to further emphasize the mediating role of appraisals, we replicated the previous experiment and had agents, instead of displaying emotions in the face, convey the appraisals directly through text. In line with reverse appraisal, this new manipulation led to virtually identical effects on people's perception of how likely agents were to cooperate in the future.

3.1 Study 1: The Effect of Emotion Displays

3.1.1 Motivation

The goal of the first study (de Melo, Carnevale, & Gratch, 2010; de Melo, Carnevale, & Gratch, 2012) was to show that people's decision to cooperate in a social dilemma can be influenced by computer agents' appropriate emotion displays. To accomplish this, a repeated-measures experiment was conducted where participants played 25 rounds of the iterated prisoner's dilemma with two different computational agents for a chance to win real money. The agents followed the same strategy to choose their actions but, showed different emotion facial displays according to the outcome of each round. The expressively cooperative agent's displays reflected a goal of reaching mutual cooperation. The expressively competitive agent's displays reflected a goal of maximizing its own points. We expected that people's decision-making would be influenced by the differences in the patterns of facial displays and hypothesized that: People would cooperate more with the cooperative than the competitive agent (H1.1).

3.1.2 Order of Play

One mediating factor that is inherent to experimental paradigms where participants play in sequence with two or more counterparts is order of play. Tversky and Kahneman (1979, 1981) show evidence that people's decision-making is reference-dependent, i.e., outcomes are expressed as positive or negative deviations (gains or losses) from a neutral reference outcome, which is assigned a value of zero. In line with this cognitive heuristic, the order with which participants play the agents is likely to influence cooperation rates. Effectively, it has been shown before that ordering effects occur when people play the iterated prisoner's dilemma in sequence with a cooperator and a non-cooperator. Harford and Solomon (1967) found that a "reformed sinner" strategy (a change in behavior from less cooperation to more cooperation) elicited higher levels of cooperation than other strategies, such as the "pacifist" strategy (where the opponent is cooperative from start to end). Also, Bixenstine and Wilson (1963) found that initial non-cooperation followed by cooperative behavior elicited higher levels of cooperation. These effects are similar to the well-studied contrast effect in the negotiation literature known as the *black-hat/white-hat* (or *bad-cop/good-cop*) effect (Hilty & Carnevale, 1993). Effectively, Hilty and Carnevale (1993) showed that playing a first game with an opponent with a competitive stance (black-hat) followed by a second game with an opponent with a cooperative stance (white-hat) was more effective in reducing distance to agreement than any other pairing of the black-hat and white-hat opponents (white-hat/white-hat, white-hat/black-hat and black-hat/black-hat). One explanation for the effectiveness of the black-hat/white-hat strategy relies on the dynamics of reciprocity. Reciprocity in negotiation is manifest in "matching" or strategy imitation, in which a bargainer concedes when the other concedes, or is firm when the other is perceived as firm (Pruitt & Carnevale, 1993). Whether people will match concessions, is dependent on context: if concessions are attributed to weakness, this will encourage exploitation

(Deutsch, Epstein, Canavan, & Gumpert, 1967). This suggests that initial firmness may lessen the temptation to exploit and that cooperative initiatives that are extended in the context of firmness may be more likely to evoke reciprocity. Another explanation of the black-hat/white-hat effect is based on the concepts of adaptation and comparison level (Helson, 1964). Theories of adaptation propose that people become accustomed to a neutral reference point as a result of prior experience; this point then serves as a comparison for judgment of subsequent experiences. (This view is closely related to Tversky and Kahneman's reference-dependency heuristic.) Thus, a cooperative second bargainer should be judged as more cooperative if the first bargainer was competitive rather than cooperative. This positive shift in perception of cooperativeness should, in turn, foster mutual cooperation. The contrast effects described here are relevant for this experiment as people always played with two different agents and, thus, we expected order of play to impact people's decision making. Following the predictions of the white-hat/black-hat effect, we hypothesized that: People would cooperate more with the cooperative agent, predominantly after playing the competitive agent first (H1.2).

3.1.3 Method

Game. Following the approach by Kiesler, Waters and Sproull (1996), the prisoner's dilemma game was recast as an investment game and described as follows to the participants: "You are going to play a two-player investment game. You can invest in one of two projects: Project Green and Project Blue. However, how many points you get is contingent on which project the other player invests in. So, if you both invest in Project Green, then each gets 5 points. If you choose Project Green but the other player chooses Project Blue, then you get 3 and the other player gets 7 points. If, on the other hand, you choose Project Blue and the other player chooses Project Green, then you get 7 and the other player gets 3 points. A fourth possibility is that you both choose Project Blue, in which case both get 4 points". There were, therefore, two possible actions in each

round: *Project Green* (or cooperation); and *Project Blue* (or defection). Table 3.1 summarizes the payoff matrix. The participant was told that there was no communication between the players before choosing an action. Moreover, the participant was told that the agent made its decision without knowledge of what the participant's choice in that round was. *After* the round was over, the action each chose was made available to both players and the outcome of the round, i.e., the number of points each player got, was also shown. The participant was informed it would play 25 rounds of the investment game with two different parties. The experiment was fully implemented in software and a snapshot is shown in Figure 3.1: During game play, the payoff matrix was shown on the top right, the outcome of the previous round in the upper mid right, the total outcome and the actions in the previous round in the lower mid right, the possible actions on the bottom right and the real-time animation of the agent on the left.

Table 3.1. Payoff matrix for the prisoner's dilemma in Study 1.

		Agent			
		Project Green		Project Blue	
Participant	Project Green	Participant:	5 pts	Participant:	3 pts
		Agent:	5 pts	Agent:	7 pts
	Project Blue	Participant:	7 pts	Participant:	4 pts
		Agent:	3 pts	Agent:	4 pts

Action Policy. Agents in both conditions played the same action policy, i.e., they followed the same strategy to choose their actions. The policy was a variant of *tit-for-tat*. Tit-for-tat is a strategy where a player begins by cooperating and then proceeds to repeat the action the other player did in the previous round. Tit-for-tat has been argued to strike the right balance of punishment and reward with respect to the opponent's previous actions (Axelrod, 1984). So, the action policy used in our experiment was as follows: (a) in rounds 1 to 5, the agent played the following fixed sequence: cooperation, cooperation, defection, defection, cooperation; (b) in

rounds 6 to 25, the agent played pure tit-for-tat. The rationale for the sequence in the first five rounds was to make it harder for participants to learn the agents' strategy and to allow participants to experience a variety of facial displays from the start.

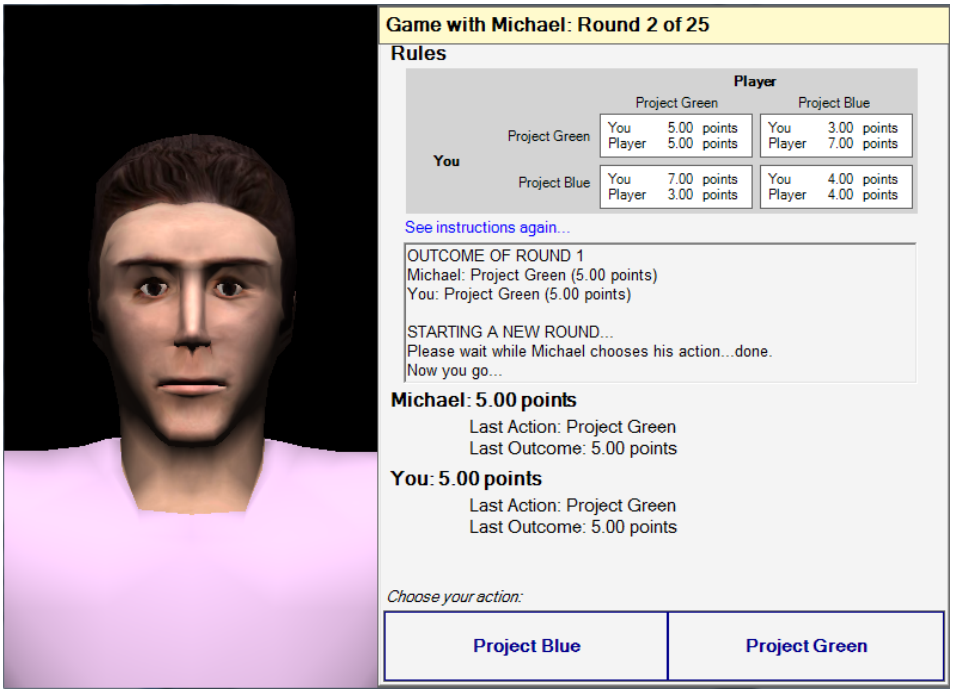


Figure 3.1. The software used in Study 1.

Conditions. There were two conditions in this experiment: the expressively cooperative agent; and the expressively competitive agent. Both agents followed the same action policy but differed in their facial display policies. The facial display policy defines the emotion and intensity which is conveyed for each possible outcome of a round. Table 3.2 shows the facial displays for the cooperative agent and Table 3.3 for the competitive agent. The facial displays were chosen to reflect the agents' goals in a way that was consistent with appraisal models of emotion (Ellsworth & Scherer, 2003). The cooperative agent's displays reflected a goal of reaching mutual cooperation. Thus, when both players cooperated, it expressed joy, as the outcome was appraised

to be positive for both players; when the agent defected and the participant cooperated, it expressed guilt, as the outcome was negative for the participant and the agent was responsible; when the agent cooperated and the participant defected, it expressed anger, as the outcome was negative and the participant was blamed for it; and, when both defected, it expressed sadness, as the event was negative. The competitive agent's displays, on the other hand, and in line with the definition of a competitive social value orientation (McClintock & Liebrand, 1988), reflected a goal of maximizing its own points. Therefore, when the agent defected and the participant cooperated, it expressed joy, as this event was appraised to be very positive; when both cooperated, it expressed nothing, as this event could be more positive; when both defected, it expressed sadness at 50%, as this was a negative event; when the participant defected and the agent cooperated, it expressed sadness at 100%, as this was the worst event for the self. Facial displays were animated using a real-time pseudo-muscular model for the face which also simulated wrinkles and blushing (de Melo & Gratch, 2009a; de Melo & Paiva, 2006a; for a detailed overview of the virtual humans platform see the Appendix). The facial display was shown at the end of the round, after both players had chosen their actions and the outcome was shown. Moreover, there was a 4.5 seconds waiting period before the participant was allowed to choose the action for the next round. This period allowed the participant to appreciate the outcome of a round before moving to the next round. Finally, to enhance naturalness, blinking was simulated in both agents as well as subtle random motion of the neck and back.

Table 3.2. Facial displays for the expressively cooperative agent (Study 1).

Expressively Cooperative		*Agent*	
		Project Green	Project Blue
Participant	Project Green	Joy (100%)	Guilt (100%)
	Project Blue	Anger (100%)	Sadness (100%)

Table 3.3. Facial displays for the expressively competitive agent (Study 1).

Expressively Competitive		*Agent*	
		Project Green	Project Blue
Participant	Project Green	Neutral	Joy (100%)
	Project Blue	Sadness (100%)	Sadness (50%)

The condition order was randomized while making sure that 50% of the participants experienced one order and the remaining 50% the other. Two different bodies were used: Michael and Daniel. These bodies are shown in Figure 3.2 as well as their respective facial displays. Notice guilt was distinguished from sadness by blushing of the cheeks. Bodies were assigned to each condition in random order and agents were referred to by the names of their bodies throughout the experiment.

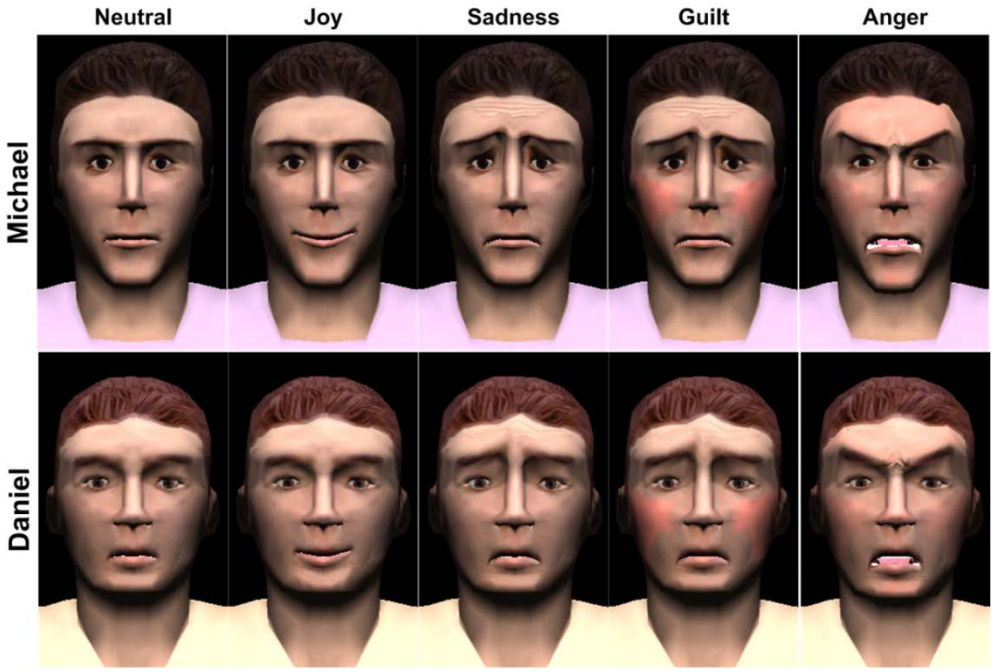


Figure 3.2. The agent bodies in Study 1–Michael and Daniel–and their facial displays.

To validate the facial displays, a pre-study was conducted where participants were asked to classify, from 1 (meaning 'not at all') to 5 (meaning 'very much'), how much each of the displays conveyed joy, sadness, guilt and anger. Images of the displays and questions were presented in random order. Twenty-two participants were recruited just for this study from the same participant pool as the main experiment (described below). The results are shown in Table 3.4. A repeated-measures ANOVA was used to compare the means for perceived emotion in each display. Significant differences were found for all displays except, as expected, for the neutral case. Moreover, pairwise comparisons of the perception of the real emotion with respect to perception of the other emotions were all significant in favor of the real emotion, with one exception: displays of guilt were also significantly perceived as displays of sadness. This is not a problem since it is usually agreed that guilt occurs upon the occurrence of a negative event, thus causing sadness, plus the attribution of blame to the self (Ortony et al., 1988).

Table 3.4. Perceived emotions in the agents' facial displays (Study 1).

	Perceived Emotion			
	Joy	*Sadness*	*Guilt*	*Anger*
Real Emotion	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
		Michael		
Neutral	1.86 (0.941)	1.86 (1.037)	1.91 (1.065)	1.68 (0.945)
Joy*	4.05 (0.899)	1.18 (0.501)	1.23 (0.528)	1.41 (1.098)
Sadness*	1.27 (0.703)	4.09 (1.019)	2.77 (1.478)	1.50 (0.859)
Guilt*	1.32 (0.716)	3.59 (1.182)	3.55 (1.371)	1.45 (0.858)
Anger*	1.36 (0.727)	1.95 (1.046)	1.32 (0.646)	4.32 (1.211)
		Daniel		
Neutral	1.55 (1.057)	1.73 (0.935)	1.68 (0.894)	2.18 (1.259)
Joy*	3.77 (1.020)	1.18 (0.501)	1.23 (0.528)	1.14 (0.468)
Sadness*	1.41 (0.854)	3.68 (1.492)	2.73 (1.386)	1.50 (0.740)
Guilt*	1.32 (0.780)	3.77 (1.412)	3.86 (1.356)	1.41 (0.734)
Anger*	1.27 (0.703)	1.82 (1.332)	1.55 (1.011)	4.27 (1.420)

Note. Scale goes from 1 (meaning 'not at all') to 5 (meaning 'very much').

* Significant difference between means in same row using repeated-measures ANOVA, $p < .05$.

Measure. The main (behavioral) measure in this experiment was cooperation rate over all rounds, i.e., the number of times the participant cooperated over the number of rounds.

Participants. Fifty-one participants were recruited at the USC Marshall School of Business. Average age was 21.0 years. Gender distribution was as follows: *males*, 45.1%; *females*, 54.9%. Most participants were undergraduate students (96.9%) majoring in business (86.3%). Most were also originally from the United States (84.3%). The incentive to participate followed standard practice in experimental economics (Hertwig & Ortmann, 2001): first, participants were given school credit for their participation in this experiment; second, with respect to their goal in the game, participants were instructed to earn as many points as possible, as the total amount of points would increase their chances of winning a lottery of $100.

3.1.4 Results

To understand how people cooperated with the agents in each condition, the following variables were defined:

- Coop.All, cooperation rate over all rounds;
- Coop.AgC, cooperation rate when the agent cooperated in the previous round;
- Coop.AgD, cooperation rate when the agent defected in the previous round.

The Kolmogorov-Smirnov test was applied to all these variables to test for their normality and all were found to be significantly non-normal. Therefore, the Wilcoxon signed ranks test was used to compare means between conditions. The results, shown in Table 3.5, indicated that people cooperated significantly more with the cooperative agent (M = .37, SD = .28) than the competitive agent (M = .27, SD = .23; $p < .05$, r = .320). Thus, our hypothesis H1.1 was confirmed. The results also suggested that this difference in cooperation was particularly salient following a defection by the agent.

Table 3.5. Descriptive statistics and significance for cooperation rate (Study 1).

Variables	*Cooperative*		*Competitive*		*Sig.*	*|r|*
	Mean	*SD*	*Mean*	*SD*	*2-sd*	
Coop.All*	.366	.279	.272	.231	.022	.320
Coop.AgC	.397	.319	.339	.288	.262	ns
Coop.AgD*	.297	.256	.203	.197	.022	.320

* $p < .05$.

Figure 3.3 shows how cooperation rate (Coop.All) evolved with each round. The graph shows that people started cooperating less with the competitive agent as early as the 3rd round. Even though both agents defected in rounds 3 and 4 (see 'Action Policy' subsection), participants cooperated much less with the competitive agent in round 5. After the agents cooperated in rounds 5 and 6, people seemed to attempt cooperation again in round 7 with the competitive agent but, from then on, again consistently cooperated less with the competitive agent.

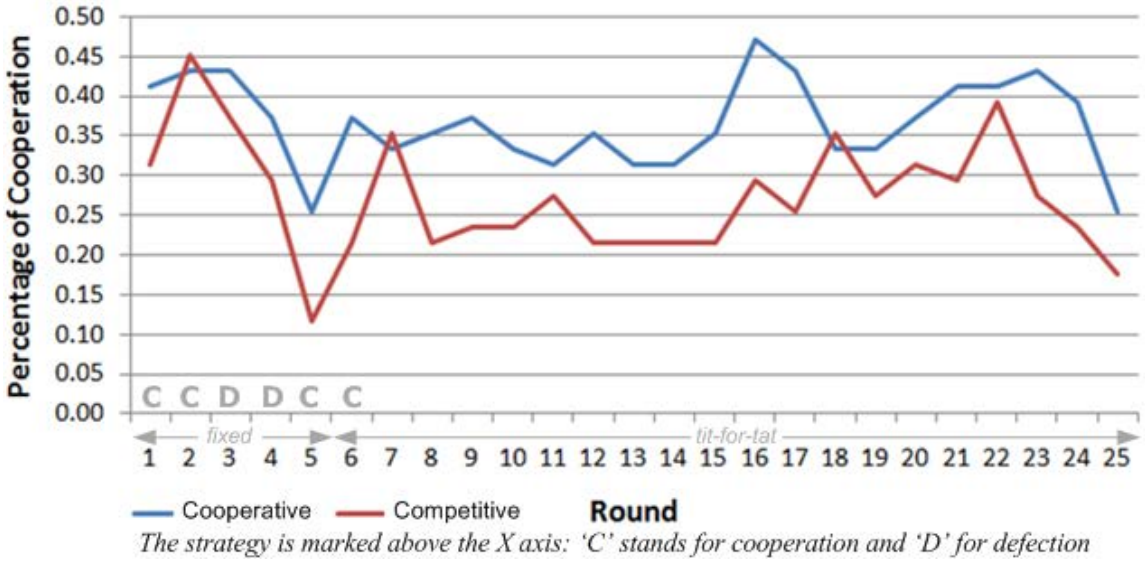


Figure 3.3. Cooperation rate per round (Study 1).

As discussed above, studies have shown that when people engage in sequence with a cooperator and a non-cooperator in a social dilemma, the order of interaction can have an impact on level of cooperation (Bixenstine & Wilson, 1963; Harford & Solomon, 1967; Hilty & Carnevale, 1993). To explore whether order had an effect in cooperation, Table 3.6 shows cooperation rates for each condition order. The results were clear and revealed that the effect described above (Table 3.5) was driven by the order competitive agent first, cooperative agent second. Effectively, cooperation did not differ significantly between conditions when participants played with the cooperative agent first. Therefore, our hypothesis H1.2 was also confirmed.

Table 3.6. Cooperation rates by condition order (Study 1)

Variables	*Cooperative*		*Competitive*		*Sig.*	*|r|*
	Mean	*SD*	*Mean*	*SD*	*2-sd*	
	Cooperative → Competitive (n=26)					
Coop.All	.345	.260	.309	.261	.572	ns
Coop.AgC	.380	.301	.367	.314	.897	ns
Coop.AgD	.267	.203	.207	.203	.232	ns
	Competitive → Cooperative (n=25)					
Coop.All*	.389	.302	.234	.192	.016	.484
Coop.AgC	.414	.342	.310	.260	.159	ns
Coop.AgD*	.329	.303	.199	.195	.064	.370

* $p < .05$.

Since there is evidence that people form judgments of people based only on appearance (Willis & Todorov, 2006), we wanted to make sure that the body was not a confounding factor in our experiment. Thus, we compared percentage of cooperation between the two agent bodies used in the experiment. It was found that there was no significant difference in cooperation between Michael (M = .33, SD = .26) and Daniel (M = .31, SD = .26; $p > .05$). Significance was calculated using the Wilcoxon signed ranks test.

3.1.5 General Discussion

The results showed that people cooperated more with the expressively cooperative agent than with the expressively competitive agent (hypothesis H1.1). The results were, thus, in line with predictions that nonverbal behavior can impact emergence of cooperation (Boone & Buck, 2003; Frank, 1988; Nesse, 1990; Trivers, 1979; Van Kleef et al., 2010) and with the social-functions view of emotion which argues emotions convey information about one's intentions in social encounters (Frijda & Mesquita, 1994; Keltner & Haidt, 1999; Keltner & Kring, 1998; Morris & Keltner, 2000; Oatley & Jenkins, 1996). Moreover, because the displays were carefully chosen to reflect appraisals that are consistent with specific social value orientations, the results also suggest that people infer about the agents' propensity for cooperation by retrieving information, from the emotion displays, about how the agents are appraising the social dilemma outcomes.

The results also revealed an order effect: People only cooperated significantly more with the cooperative agent after playing first the competitive agent (hypothesis H1.2). This is in line with the well-studied contrast effect known as the black-hat/white-hat effect (Bixenstine & Wilson, 1963; Harford & Solomon, 1967; Hilty & Carnevale, 1993). When applied to our study, this meant the cooperative agent was interpreted as the white-hat and the competitive agent as the black-hat. However, whereas in the classical studies a cooperative or competitive stance was signaled through different levels of concession in the offers, in the current study this was signaled through emotion displays. The argument, then, is that when participants face a tough competitive agent in the first game, they will be less likely to attempt exploitation in the second game and reciprocate to a more (expressively) cooperative agent. Effectively, the results suggested that participants exploited the cooperative agent less, after playing the competitive agent first.

3.2 Study 2: The Importance of Context

The second study (de Melo, Carnevale, & Gratch, 2011b) was designed to further unpack the mechanism for the interpersonal effect of emotion in social dilemmas. We looked at appraisal theories (Ellsworth & Scherer, 2003) for an explanation of how participants came to understand the meaning behind the emotion displays in the first study. According to this appraisal perspective, the impact of an emotional display depends on the inferences about mental state that people make from the information in the displays; furthermore, these inferences require an interpretation of the expression in the context of the events that led to its production. Thus, the identical emotional display (e.g., a smile) might have very different meaning depending on if it is produced in response to a cooperative or competitive action. Effectively, Hareli and Hess (2010) had already noticed that the same smile can mean different things about someone's character according to the context in which is shown. Thus, in this study, we explored a design that emphasized the importance of context for the interpretation of emotion displays and proposed two new versions of the cooperative and competitive agents that only differed in the context in which joy was expressed. The expressively cooperative agent (Table 3.7) smiled when mutual cooperation occurred, whereas the expressively competitive agent (Table 3.8) smiled when it exploited the participant. Both agents showed anger when the participant exploited the agent and, otherwise, showed no emotion. The new design would also allow us to exclude two alternative explanations to the results in the first study: (1) people cooperated more with the cooperative agent because it showed more emotion than the competitive agent, since the competitive agent did not show emotion in mutual cooperation; (2) people cooperated more with the cooperative agent because it showed more emotion types (joy, sadness, anger and guilt) than the competitive agent (joy and sadness).

Table 3.7. Facial displays for the expressively cooperative agent (Study 2).

Expressively Cooperative		*Agent*	
		Project Green	Project Blue
Participant	Project Green	Joy (100%)	Neutral
	Project Blue	Anger (100%)	Neutral

Table 3.8. Facial displays for the expressively competitive agent (Study 2).

Expressively Competitive		*Agent*	
		Project Green	Project Blue
Participant	Project Green	Neutral	Joy (100%)
	Project Blue	Anger (100%)	Neutral

This study also addressed a limitation in the first study: there was no comparison to a control no-emotion agent. Thus, in the previous study it wasn't clear whether people were cooperating more with the cooperative agent or less with the competitive agent. Therefore, this study consisted of three experiments: (1) cooperative vs. control; (2) competitive vs. control; (3) cooperative vs. competitive. Because previous studies in human-machine interaction have shown that people tend to prefer emotional over non-emotional agents (de Melo et al., 2009; Hone, 2006; Klein et al., 2002; Lester et al., 1997; Lim & Aylett, 2007; Liu & Picard, 2005; Maldonado et al., 2005; Prendinger et al., 2003), our expectation was that the control agent would be perceived as a non-cooperator. The control agent's lack of emotional responsiveness was likely to lead participants to perceive it as being a tough opponent. In contrast, the competitive agent was expected to be perceived as a non-cooperator due to its selfish displays. Because the cooperative agent should be perceived as a cooperator (white-hat) and the control and competitive agents as non-cooperators (black-hats), we expected ordering effects to occur in this study as well (see Subsection 3.1.2). Therefore, we advanced the following hypotheses for this study: regarding the cooperative vs. control experiment, people would cooperate more with the cooperative agent

(H2.1) and, because of the black-hat/white-hat contrast effect, the effect would be driven by the order control → cooperative agent (H2.2); regarding the competitive vs. control experiment, in line with expectations for the black-hat/black-hat pattern (Hilty & Carnevale, 1993), people would not cooperate differently with the competitive and control agents (H2.3); finally, regarding the cooperative vs. competitive experiment, people would cooperate more with the cooperative agent (H2.4) and, because of the black-hat/white-hat contrast effect, the effect would be driven by the order competitive → cooperative agent (H2.5).

3.2.1 Experiment 1: Cooperative vs. Control

3.2.1.1 Method

The game used in this experiment was the same as in the first study, except that the payoff matrix was slightly modified, reducing the payoff for the exploited player from 2 to 3, as shown in Table 3.9. The game software interface was also updated, as shown in Figure 3.4. The fixed-sequence part of the action policy (first five rounds) was changed to: cooperation, cooperation, defection, *cooperation*, cooperation. This avoided the double-defection pattern in the original sequence. Regarding conditions, the participant played with the cooperative and the no-emotion control agents. Two new agent bodies were used, Figure 3.5: Ethan and William. Finally regarding measures, as in the previous study, we measured cooperation rate over all rounds.

Table 3.9. Payoff matrix for the prisoner's dilemma in Study 2.

		Agent	
		Project Green	Project Blue
Participant	Project Green	Participant: 5 pts Agent: 5 pts	Participant: 2 pts Agent: 7 pts
	Project Blue	Participant: 7 pts Agent: 2 pts	Participant: 4 pts Agent: 4 pts

Figure 3.4. The software used in Study 2.

Forty-eight participants were recruited at the USC Marshall School of Business. Average age was 21.6 years and 62.5% were males. Most participants were undergraduate (41.7%) or graduate (56.3%) students majoring in diverse fields. Most were originally from Asia (66.7%) and North America (33.3%). The incentive to participate followed standard practice in experimental economics (Hertwig & Ortmann, 2001): first, participants were given $15 for their participation in this experiment; second, with respect to their goal in the game, participants were instructed to earn as many points as possible, as the total amount of points would increase their chance of winning a lottery for $100.

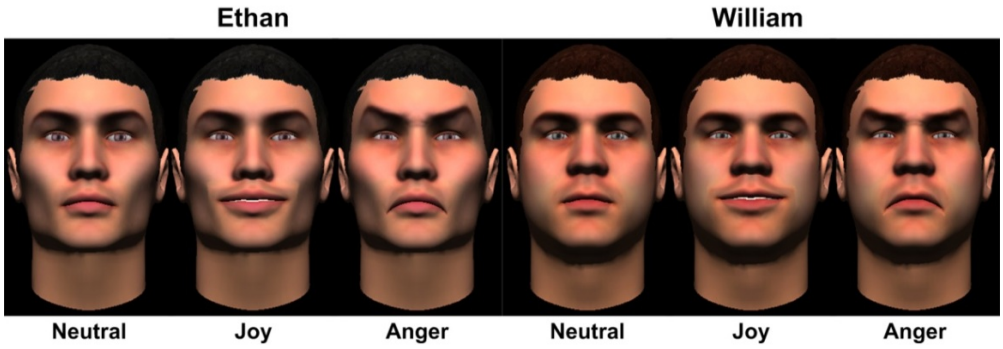


Figure 3.5. The agent bodies in Study 2–Ethan and William–and their facial displays.

3.2.1.2 Results

Participants that did not experience joy with the cooperative agent[2] were excluded from analysis (though keeping them would lead to the same pattern of results). So, 10 (out of 48) participants were excluded. To understand how people cooperated with the agents, we looked at cooperation rate over all rounds. Figure 3.6 and Table 3.10 show the results for this variable. Significance levels were calculated using the repeated-measures *t* test. The results showed that people cooperated significantly more with the cooperative agent than the control agent (Table 3.10). Thus, hypothesis H2.1 was confirmed. The results also showed that this effect was driven by the condition order in which participants played the control agent first, and the cooperative agent second (thus, confirming hypothesis H2.2).

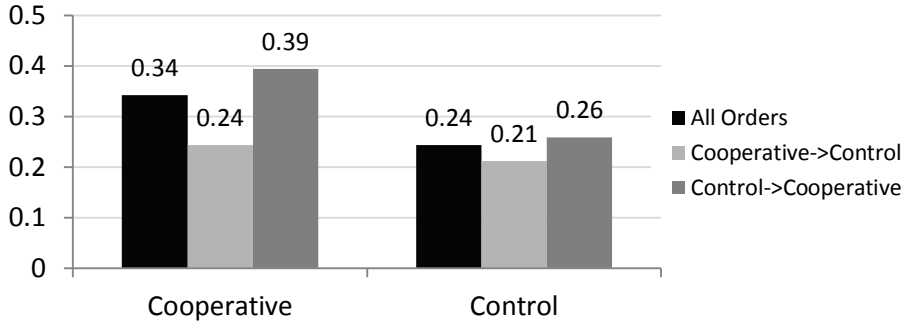


Figure 3.6. Cooperation rates in Experiment 1 (Study 2).

[2] Notice our paradigm did not guarantee participants would experience all outcomes in the prisoner's dilemma game.

Table 3.10. Cooperation rate by condition order (Study 2, Experiment 1).

Variables	*Cooperative*		*Control*		*Sig.*
	Mean	*SD*	*Mean*	*SD*	*2-sd*
		Both orders (n=38)			
Coop.All*	.342	.173	.243	.141	.006
Coop.AgC	.327	.243	.259	.181	.093
Coop.AgD*	.344	.196	.242	.197	.009
		Cooperative → Control (n=13)			
Coop.All	.243	.089	.212	.119	ns
Coop.AgC	.185	.175	.218	.174	ns
Coop.AgD	.282	.155	.225	.181	ns
		Control → Cooperative (n=25)			
Coop.All*	.394	.185	.259	.152	.008
Coop.AgC*	.402	.243	.280	.184	.038
Coop.AgD*	.376	.210	.251	.207	.018

* $p < .05$.

3.2.2 Experiment 2: Competitive vs. Control

3.2.2.1 Method

Experiment 2 followed the same method as Experiment 1, except that participants played with the competitive agent and the control agent. Thirty-eight participants were recruited from the USC Marshall School of Business. Average age was 22.3 years and 63.3% were males. Most participants were undergraduate (46.7%) or graduate (53.3%) students majoring in diverse fields. Most were also originally from Asia (66.7%) and North America (33.3%).

3.2.2.2 Results

Participants that did not experience joy with the competitive agent were excluded from analysis (though keeping them would lead to the same pattern of results). So, 8 (out of 38) participants were excluded. Figure 3.7 and Table 3.11 show the cooperation rates. The results showed that people were not cooperating differently with the expressively competitive or control agents, thus confirming hypothesis H2.3.

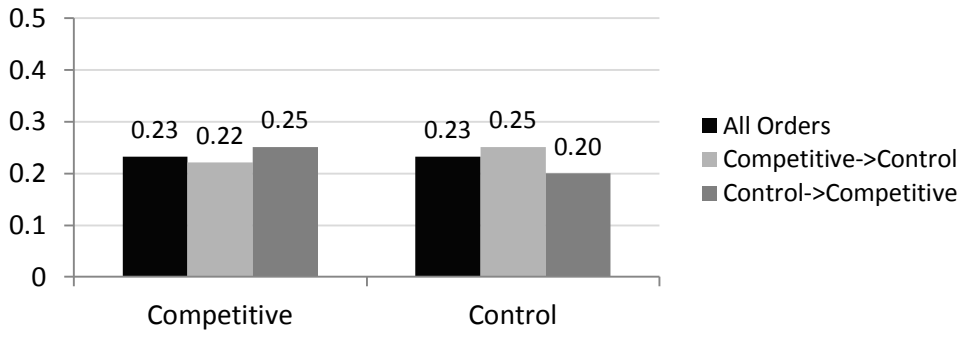


Figure 3.7. Cooperation rates in Experiment 2 (Study 2).

Table 3.11. Cooperation rate by condition order (Study 2, Experiment 2).

Variables	*Competitive*		*Control*		*Sig.*
	Mean	*SD*	*Mean*	*SD*	*2-sd*
	Both orders (n=30)				
Coop.All	.232	.109	.232	.170	ns
Coop.AgC	.225	.195	.246	.194	ns
Coop.AgD	.236	.111	.222	.177	ns
	Competitive → Control (n=19)				
Coop.All	.221	.096	.251	.178	ns
Coop.AgC	.219	.202	.272	.215	ns
Coop.AgD	.225	.115	.235	.184	ns
	Control → Competitive (n=11)				
Coop.All	.251	.133	.200	.158	ns
Coop.AgC	.236	.191	.200	.149	ns
Coop.AgD	.256	.108	.199	.170	ns

* $p < .05$.

3.2.3 Experiment 3: Cooperative vs. Competitive

3.2.3.1 Method

Experiment 3 followed the same method as Experiment 1, except that participants played with the cooperative and competitive agents. Fifty-one participants were recruited from the USC Marshall School of Business. Average age was 22.0 years and 62.7% were males. Most participants were

undergraduate (54.9%) or graduate (43.2%) students majoring in diverse fields. Most were also originally from Asia (52.9%) and North America (47.1%).

3.2.3.2 Results

Participants that did not experience joy at least once with *each* agent were excluded from analysis. So, 13 (out of 51) participants were excluded. Figure 3.8 and Table 3.12 shows the cooperation rates. The results showed that, for the order competitive → cooperative, people cooperated significantly more with the cooperative agent thus, confirming hypothesis H2.5. However, when collapsing across orders, there was no significant difference in cooperation between the agents and, thus, hypothesis H2.4 could not be confirmed.

Table 3.12. Cooperation rate by condition order (Study 2, Experiment 3).

Variables	*Cooperative*		*Competitive*		*Sig.*
	Mean	*SD*	*Mean*	*SD*	*2-sd*
	Both orders (n=34)				
Coop.All	.413	.225	.393	.211	ns
Coop.AgC	.404	.288	.403	.280	ns
Coop.AgD	.393	.269	.364	.249	ns
	Cooperative → Competitive (n=14)				
Coop.All	.374	.183	.494	.193	.120
Coop.AgC	.382	.237	.527	.272	.098
Coop.AgD	.366	.229	.418	.262	.616
	Competitive → Cooperative (n=20)				
Coop.All*	.440	.252	.322	.198	.044
Coop.AgC	.420	.324	.317	.258	.151
Coop.AgD	.411	.298	.326	.239	.239

* $p < .05$.

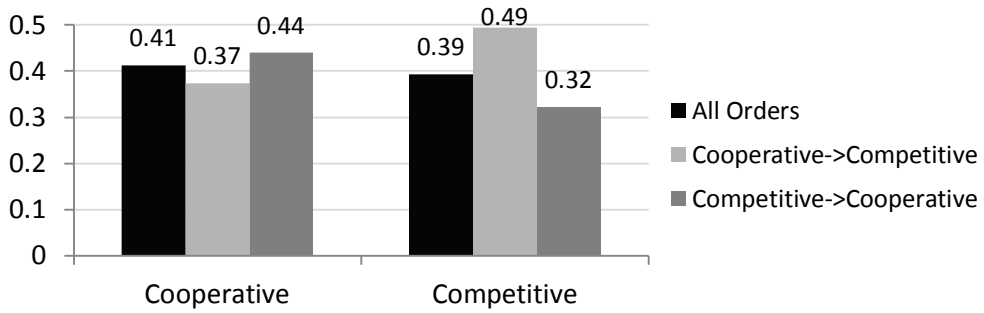


Figure 3.8. Cooperation rates in Experiment 3 (Study 2).

3.2.4 General Discussion

Experiment 1 showed that people cooperated more with the cooperative agent than the no-emotion control agent (hypothesis H2.1) and, that the effect was driven by the order control → cooperative (hypothesis H2.2). Experiment 2 showed that people did not cooperate differently between the competitive and control agents (hypothesis H2.3). Together, these results suggest the cooperative agent was perceived as a cooperator (white-hat) and the competitive and control agents as non-cooperators (black-hats). Thus, in Experiment 3, we expected people to cooperate more with the cooperative agent, especially in the black-hat/white-hat order. The results showed, effectively, that when playing with the competitive agent first (black-hat/white-hat order), people cooperated significantly more with the cooperative agent (hypothesis H2.5). However, when collapsing across orders, there was no significant difference in cooperation rates between the cooperative and competitive agents. Hypothesis H2.4 was, thus, not confirmed. Moreover, when playing with the cooperative agent first (white-hat/black-hat order), there was an unexpected trend to cooperate more with the competitive agent. One possible explanation for this is based on adaptation level theory (Helson, 1964) which predicts high concessions in response to the white-hat/black-hat sequence because the black hat will appear toughest when preceded by a white hat; in a competitive context, this enhances the tendency to yield to a powerful opponent. A negative

shift in cooperation can also evoke more concessions if it produces a desire to entice the black-hat adversary with cooperative gestures to return to former levels of cooperation. Finally, Hilty & Carnevale (1993) also report that, in bilateral negotiation, a negative shift in cooperation, elicits "unilateral concessions from participants in an effort to induce the bargainer to resume former levels of concession-making" (p. 458).

Overall, the results emphasize the importance of context for interpreting emotion displays. Effectively, the expressively cooperative and expressively competitive agents only differed in the context in which they expressed joy and, yet, people played completely differently with each. These results are in line with appraisal theories, which argue that interpreting emotions requires understanding the circumstances, or context, that led to the generation of the emotion (Ellsworth & Scherer, 2003), as well as with other findings that suggest the meaning of a smile varies according to context (Hareli & Hess, 2010; Van Kleef et al., 2010).

3.3 Study 3: Reverse Appraisal

To provide direct evidence for reverse appraisal, a third study was conducted to examine the contextual effects of emotion reported in the previous study. Following findings of Hareli and Hess (2010), people were hypothesized to use emotional displays to infer beliefs, desires and intentions of their social partners essentially by reversing the appraisal mechanism. The focus in the new study was, thus, on the following questions: what was the information conveyed by emotion displays? How was this information retrieved from emotion displays? How did this information influence beliefs relevant to decision-making? According to reverse appraisal, people infer, from emotion displays, how the counterpart is appraising the social dilemma outcomes; secondly, from these *perceptions of appraisal*, people infer how likely the counterpart is to cooperate in the future, which we refer to as *perceptions of cooperativeness*. This causal model is

shown in Figure 3.9. The goal in this study was to establish this model[3]. To accomplish this, two new experiments are presented: the first demonstrates mediation of perceptions of appraisal on the effect of emotion displays on perceptions of cooperativeness using a statistical method, as suggested by Baron and Kenny (1986); the second, following a suggestion by Spencer, Zanna and Fong (2005), establishes the model experimentally.

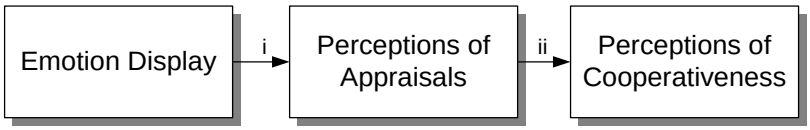


Figure 3.9. Proposed causal model for the impact of emotion displays in decision-making.

3.3.1 Experiment 1: Appraisal Mediation

In Experiment 1, participants were given scenarios where they imagined playing the iterated prisoner's dilemma with agents that displayed emotion. Each scenario pertained to the first round (of a 5-round game) and corresponded to a particular outcome of the game. Participants were then shown a video of how the other side reacted to the outcome. The reaction corresponded to a facial display of emotion. Participants were then asked to assess (a) the emotion being displayed (b) how the other player was appraising the outcome (perceptions of appraisal) and, (c) how likely the other player was to cooperate in the future (perceptions of cooperativeness). In line with reverse appraisal, we hypothesized that, for a given outcome, different emotion displays would lead to different patterns of perceptions of appraisal, in a way that was consistent with expectations from appraisal theories (H3.1). Additionally, we hypothesized that the emotion display manipulation would influence perception of the agent's cooperativeness (H3.2). A multiple mediation analysis (Preacher & Hayes, 2008) was also conducted to understand, for each

[3] Notice that showing one perceives the other is a cooperator is not the same as saying that the former will cooperate. The causal link between perception of the other side's cooperativeness and the actual decision to cooperate is not simple–e.g., whereas a pro-social might cooperate, a pro-self might exploit the cooperator (Steinel & de Dreu, 2004). See the Discussion Chapter for further details on how to address this link.

outcome, the mediating role of perceptions of appraisal on the effect of emotion displays on perceptions of cooperativeness. Our hypothesis was that perceptions of appraisal would mediate, at least partially, this effect (H3.3).

3.3.1.1 Method

Game. As in the previous studies, the prisoner's dilemma game was recast as an investment game (Kiesler et al., 1996). The scenarios pertained to the first round of the iterated prisoner's dilemma, where supposedly 5 rounds would be played.

Conditions. The experiment followed a mixed design with two factors: Outcome (between-participants) with 4 levels (one for each possible outcome of the game); and, Emotion (repeated-measures) with 5 levels (Neutral vs. Joy vs. Anger vs. Sadness vs. Guilt). The design built on the experience gained from the previous studies; therefore, first, only the 4 emotions explored in those studies were considered; second, we did not consider a full factorial design but, rather, only pairings of outcome and emotion that produced effects in these studies, as shown in Table 3.13. Considering only this subset of the possible pairings had, at least, two advantages: (1) each participant experienced at most 3 pairings (as opposed to 5 if all were considered), which constrained total participation time and, thus, reduced fatigue and boredom effects; (2) pairings that did not have a clear intuitive interpretation (e.g., displaying sadness or anger in mutual cooperation) were excluded from analysis.

Table 3.13. Pairings of outcome and emotion explored in Experiment 1 (Study 3).

		Agent	
		Project Green	Project Blue
Participant	Project Green	Neutral, Joy	Neutral, Joy, Guilt
	Project Blue	Neutral, Anger, Sadness	Neutral, Joy, Anger

Emotion Displays. In this study, participants watched videos of virtual agents expressing facial emotion displays. Three agents were used–Ethan, William and David–and the respective facial displays are shown in Figure 3.10. The agents were referred to by their names throughout the experiment. Each participant saw a different agent in each condition, and they were randomly assigned to conditions.

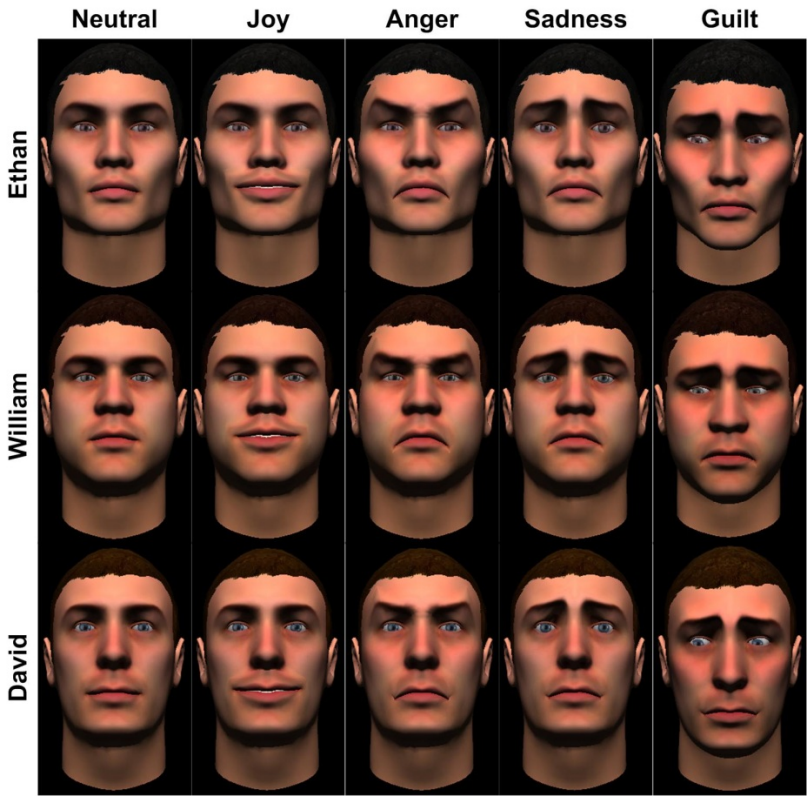


Figure 3.10. The emotion facial displays used in Experiment 1 (Study 3).

Measures for emotion interpretation. After watching the video of the agent's emotional reaction, we asked participants the following questions (the questions referred to the agents by their respective names): How much did the agent experience each of the following emotions a) Sadness b) Joy c) Anger d) Guilt? (scale goes from 1, *not at all*, to 7, *very much*).

Measures for perception of appraisals. Even though several appraisal theories have been proposed (Ellsworth & Scherer, 2003; Frijda, 1989; Ortony et al., 1988; Roseman, 2001; Scherer, 2001; Smith & Ellsworth, 1985), there tends to be agreement on which appraisals predict the emotions we consider in this experiment: joy occurs when the event is conducive to one's goals; anger occurs when the event is not conducive to one's goals, is caused by another agent and one has power/control over it; sadness occurs when the event is not conducive to one's goals; guilt occurs when the event is not conducive to one's goals and is caused by the self. Thus, three appraisal variables are of relevance here: (a) *conduciveness to goals*, which measures whether the event is consistent or inconsistent with the individual's goals; (b) *blameworthiness*, which measures whether the self or another agent is responsible for the event; (c) *coping potential*, which measures one's ability to deal with (or control) the consequences of an event. Thus, after watching the video of the agent's emotional reaction, participants were asked the following questions about how the agent was appraising the outcome:

1. How pleasant for him was it to be in this situation? (conduciveness to goals; Smith & Ellsworth, 1985)
2. At the time of experiencing the emotion, do you think he perceived that the consequences of the event did or would bring about positive, desirable consequences for him (e.g., helping him reach a goal, or giving pleasure)? (conduciveness to goals; Scherer, 2001)
3. Was the situation obstructive or conducive to his goals? (conduciveness to goals; Frijda, 1989)
4. Was what happened something that he regarded as unfair or fair? (conduciveness to goals; Frijda, 1989)
5. At the time, how much did you think he blamed himself for the event? (self-blameworthiness; Smith & Ellsworth, 1985)

6. At the time, how much did you think he blamed you for the event? (participant-blameworthiness; Smith & Ellsworth, 1985)
7. After he had a good idea of what the probable consequences of the event would be, do you think that he: (coping potential; Scherer, 2001)
    a. would be able to avoid the consequences or modify them to his advantage (through his own power)?
    b. could 'live with', and adjust to, the consequences of the event that could not be avoided or modified?
8. During the event, do you think he felt powerless or powerful? (coping potential; Roseman & Spindel, 1990)
9. At the time, do you believe he was unable to cope with the event or that he was able to cope with it? (coping potential; Roseman & Spindel, 1990)

Measure for perception of cooperativeness. Following the appraisal perception questions, we asked the participant one question about perception of the agent's cooperativeness (scale goes from 1, *not at all*, to 7, *very much*): How likely is he to choose GREEN in the next round?

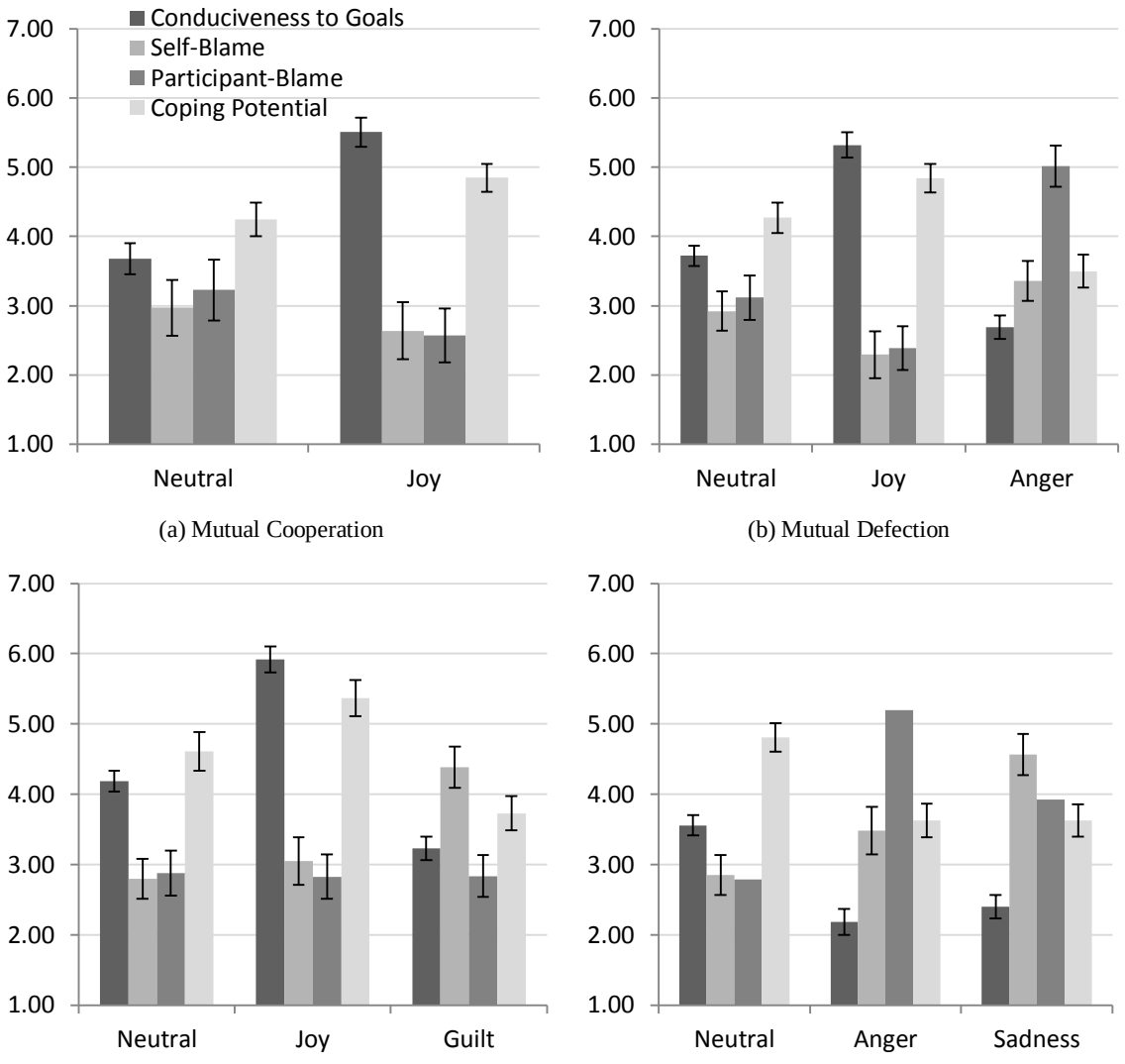
Participants. We recruited four-hundred and five (N=405) participants online using Amazon Mechanical Turk. This resulted in approximately 100 participants for each outcome. Gender distribution was as follows: *males*, 47.4%; *females*, 52.6%. Age distribution was as follows: *18 to 21 years*, 14.1%; *22 to 34 years*, 59.5%; *35 to 44 years*, 13.6%; *45 to 54 years*, 7.9%; *55 to 64 years*, 4.2%; *65 years and over*, 3.0%. Most participants were from the United States (57.8%) and India (29.6%). The education level distribution was as follows (current or expected degrees): *high school*, 15.8%; *college*, 57.5%; *Masters*, 23.0%; *Ph.D. or above*, 3.7%. Education majors and

profession were quite diverse. Participants were paid USD $1.02 and average participation time was 23 minutes.

3.3.1.2 Results

Effects for perception of appraisals. Questions 1 to 4 were highly correlated (α = .850) and, thus, were collapsed (averaged) into a single measure called *conduciveness to goals*. Questions 7 to 9 were also correlated (α = .613) and were collapsed into a single variable called *coping potential*. For each outcome, we conducted a repeated-measures ANOVA to analyze the effect of emotion display on conduciveness to goals, self-blameworthiness (question 5), participant-blameworthiness (question 6) and coping potential. Figure 3.11 shows means and standard errors. 0 reports means, standard deviations, significance levels and effect sizes. In summary: (a) in mutual cooperation, participants perceived the agent that smiled to find the outcome more conducive, blame less, and have higher coping potential than the agent that showed no emotion; (b) when the participant was exploited, participants perceived the agent that smiled to find the outcome more conducive and have higher coping potential than the neutral agent (pairwise comparisons were based on Bonferroni post-hoc comparisons, which are not shown in 0). The neutral agent, in turn, was perceived as having higher goal conduciveness and coping potential than the agent that showed guilt; (c) when the participant exploited, the agent that showed no emotion was perceived as having higher conduciveness to goals and coping potential than the angry or sad agents. The angry agent blamed the participant the most and the sad agent blamed himself the most; (d) in mutual defection, the agent that smiled was perceived as having higher conduciveness to goals and coping potential than the agent that showed no emotion. In turn, the neutral agent was perceived as having higher conduciveness to goals and coping potential than the angry agent. The angry agent was also perceived as blaming the participant the most. Overall,

the effect of emotion displays on perceptions of appraisals was consistent with expectations from appraisal theories thus, confirming hypothesis H3.1.



(a) Mutual Cooperation

(b) Mutual Defection

(c) Participant is Exploited

(d) Participant Exploits

Figure 3.11. Perception of appraisals in Experiment 1 (Study 3).

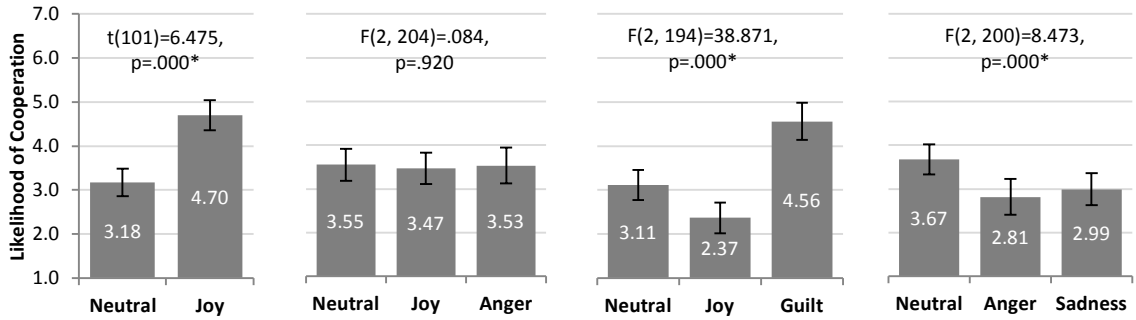
Table 3.14. Descriptive statistics for perception of appraisals in Experiment 1 (Study 3).

	Conduciveness to Goals	Blame (Self)	Blame (Participant)	Coping Potential
		Mutual cooperation (n=103)		
Neutral	3.68 (0.896)	2.97 (1.620)	3.23 (1.746)	4.25 (0.972)
Joy	5.51 (0.852)	2.64 (1.652)	2.57 (1.551)	4.86 (0.805)
Sig. (r)	.000* (.853)	.067 (.181)	.000* (.362)	.000* (.509)
		Mutual defection (n=103)		
Neutral	3.72 (0.757)	2.92 (1.453)	3.12 (1.635)	4.27 (0.867)
Joy	5.32 (0.856)	2.29 (1.493)	2.39 (1.523)	4.84 (0.831)
Anger	2.69 (0.856)	3.36 (1.726)	5.02 (1.621)	3.50 (0.950)
Sig. (η_p^2)	.000* (.733)	.000* (.119)	.000* (.477)	.000* (.462)
		Participant is exploited (n=98)		
Neutral	4.19 (1.089)	2.80 (1.699)	2.88 (1.633)	4.61 (1.093)
Joy	5.92 (0.823)	3.05 (2.078)	2.83 (1.759)	5.37 (1.024)
Guilt	3.23 (1.179)	4.39 (1.672)	2.84 (1.558)	3.73 (0.979)
Sig. (η_p^2)	.000* (.671)	.000* (.224)	.950 (.000)	.000* (.476)
		Participant exploits (n=101)		
Neutral	3.56 (1.038)	2.85 (1.676)	2.79 (1.768)	4.81 (0.800)
Anger	2.19 (0.868)	3.49 (1.659)	5.20 (1.588)	3.63 (0.954)
Sadness	2.40 (0.901)	4.56 (1.590)	3.92 (1.730)	3.63 (0.920)
Sig. (η_p^2)	.000* (.545)	.000* (.248)	.000* (.466)	.000* (.444)

* $p < .05$.

Effects for perception of cooperativeness. For each outcome, we conducted a repeated-measures ANOVA to analyze the effect of emotion display on perception of cooperativeness. Figure 3.12 shows means and standard errors. Table 3.15 reports means, standard deviations, significance levels and effect sizes. In summary: (a) in mutual cooperation, the agent that smiled was perceived as being more cooperative than the neutral agent; (b in mutual defection, there were no effects of emotion; (c) when the participant was exploited, the agent that showed guilt was perceived as more cooperative than the agent that showed no emotion (pairwise comparisons were based on Bonferroni post-hoc comparisons, which are not shown in Table 3.15). In turn, the

agent that showed no emotion was perceived as more cooperative than the agent that smiled; (d) when the participant exploited, the agent that showed no emotion was perceived as more cooperative than the agent that showed sadness or anger. Overall, the results showed that emotion displays had clear effects on perception of cooperation thus, confirming hypothesis H3.2.



(a) Mutual Cooperation (b) Mutual Defection (c) Participant is Exploited (d) Participant Exploits

Figure 3.12. Perception of cooperativeness in Experiment 1 (Study 3).

Table 3.15. Descriptive statistics for perception of cooperativeness in Experiment 1 (Study 3).

Mutual cooperation (n=103)		*Mutual defection (n=103)*	
Neutral	3.18 (1.613)	Neutral	3.55 (1.856)
Joy	4.70 (1.739)	Joy	3.47 (1.835)
-	-	Anger	3.53 (2.100)
Sig. (*r*)	.000* (.542)	Sig. (η_p^2)	.920 (.001)
Participant is Exploited (n=98)		*Participant exploits (n=101)*	
Neutral	3.11 (1.716)	Neutral	3.67 (1.715)
Joy	2.37 (1.755)	Anger	2.81 (2.077)
Guilt	4.56 (2.081)	Sadness	2.99 (1.841)
Sig. (η_p^2)	.000* (.286)	Sig. (η_p^2)	.000* (.078)

* $p < .05$.

Mediation analysis. In this section we present a causal steps approach multiple mediation analysis (Preacher & Hayes, 2008) of perceptions of appraisal on the effect of emotion displays on perception of cooperativeness. This method is an extension to multiple mediators of the single-mediation analysis proposed by Baron and Kenny (1986). Figure 3.13 summarizes the mediation model. The independent variables (IVs) were the classification questions for perception of joy, anger, sadness and guilt. The dependent variable (DV) was perception of cooperativeness. The proposed mediators were the perception of appraisal variables: conduciveness to goals, self-blame, participant-blame and coping potential. According to this approach, there is mediation by a specific mediator M_x if: (1) the path, a_x, from the IV to the mediator is significant; (2) the path, b_x, from the mediator to the DV, when controlling for the IV, is significant; (3) the indirect effect, $a_x b_x$, from the IV to the DV, when controlling for the mediator, is significantly different than zero and greater than zero by a non-trivial amount. Moreover, there is mediation of the *set* of mediators when the sum of the indirect effects of all mediators is significantly different than zero. Furthermore, there is full mediation when the direct effect, *c'*, of the IV on the DV, when controlling for all the mediators, is non-significant. Finally, in the original paper, Baron and Kenny also require the total effect, *c*, from the IV to the DV (not considering any mediators), to be significant. However, many authors advocate this path need not be significant, in the multiple mediation case, for mediation to occur (Preacher & Hayes, 2008).

Table 3.16 shows the mediation analysis. The shaded cells on the *a*, *b* and *ab* path columns represent that the causal-step requirement on the respective path has been passed. In summary, confirming hypothesis H3.3, the results showed that: (a) in mutual cooperation, conduciveness to goals partially mediated the effect of joy; (b) in mutual defection, conduciveness to goals and self-blame partially mediated the effect of joy and, self-blame and participant-blame partially mediated the effect of anger; (c) when the participant was exploited, conduciveness to goals and

self-blame fully mediated the effect of joy and, conduciveness to goals and self-blame partially mediated the effect of guilt; (d) when the participant exploited, conduciveness to goals fully mediated the effect of sadness and, conduciveness to goals fully mediated the effect of anger on cooperativeness.

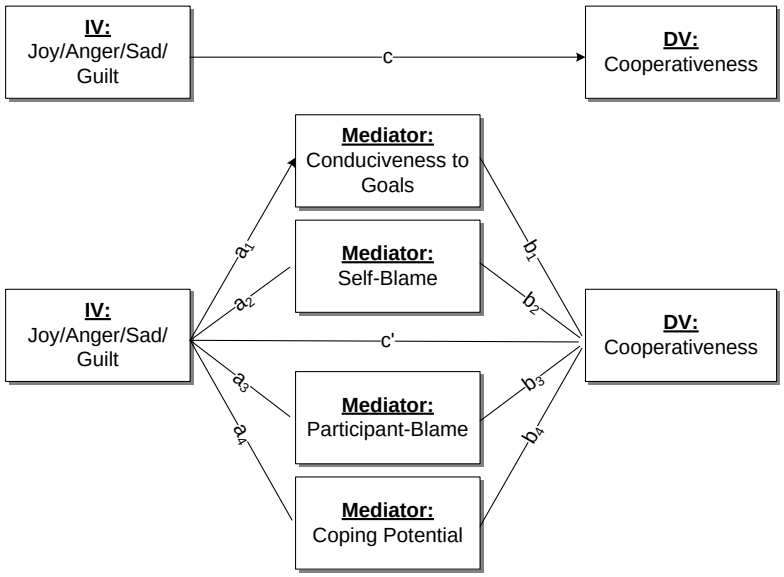


Figure 3.13. The multiple mediation model (Study 3).

3.3.1.3 Discussion

The results showed that people's perceptions about the agent's intentions were influenced by the agent's emotion displays. This is in line with findings in our previous studies and with predictions in the literature about the impact of non-verbal behavior on cooperation (Boone & Buck, 2003; Frank, 1988; Nesse, 1990; Trivers, 1979; Van Kleef et al., 2010). The experiment also showed that particular emotions impact perception of the agent's likelihood of cooperation: when the outcome was unfavorable to the agent, a display of anger or sadness signaled more unwillingness to cooperate in the future than no display of emotion; when the agent got a favorable outcome at the expense of the participant, a display of guilt showed regret and signaled willingness to cooperate in the future; a smile meant the agent was happy with the current outcome and was

likely to keep choosing the same action. The meaning of a smile was, thus, context-dependent: when the other side exploited, a smile meant unwillingness to cooperate; in mutual cooperation, a smile meant willingness to cooperate. This result is compatible with our findings in Study 2.

The results also showed that emotion displays influenced people's perception about how the agent appraised the outcomes. A smile meant the agent found the outcome conducive to his or her goals; anger meant the agent found the outcome obstructive and blamed the participant for it; sadness meant the agent found the outcome obstructive; finally, guilt meant the agent found the outcome obstructive and blamed himself for it. These patterns closely match expectations from appraisal theories (Ellsworth & Scherer, 2003; see also Section 2.5). One exception, perhaps, was coping potential. High coping potential tended to be associated with positive outcomes and low coping potential with negative outcomes. This is not necessarily in line with predictions from some appraisal theorists (e.g., Scherer, 2001) that suggest high coping potential correlates with anger. The reason for this mismatch might have been that people interpreted the questions about coping potential has referring to the ability the other side showed in the *past* to reach the current outcome, as opposed to, as suggested by Scherer, the ability of the agent to deal with the consequences of the outcome in the *future*.

Finally, the multiple mediation analysis showed that perceptions of appraisals (partially and sometimes fully) mediated the effect of emotion displays on perception of cooperativeness. Altogether, the results suggest that an important component of the information people retrieve from emotion displays pertains to how the agent is appraising the outcome. This, in turn, is in line with the reverse appraisal proposal, and constitutes evidence for the suggested causal model in Figure 3.9, i.e., emotion displays cause perceptions of appraisals, which in turn cause perceptions of cooperativeness. Experiment 2 follows a complementary approach to provide further evidence for this model.

Table 3.16. Mediation analysis of perceptions of appraisals (Experiment 1, Study 3).

		IV→ Mediators *(a paths)*				Mediators → DV *(b paths)*				Total Effect *(c path)*	Direct Effect *(c' path)*	Indirect Effect *(ab paths)*				
		Cn	SB	PB	CP	Cn	SB	PB	CP			Tot	Cn	SB	PB	CP
Mutual Cooperation	Joy	.457*	-.015	-.125*	.141*	.232	-.021	-.119	.165	.372*	.228 *	.144*	.106	.000	.015	.023
		(.000)	(.793)	(.037)	(.000)	(.124)	(.834)	(.247)	(.259)	(.000)	(.007)	(.020)	(.120)	(.869)	(.301)	(.267)
Mutual Defection	Joy	.501*	-.140*	-.360*	.220*	.281*	.255*	-.073	-.083	-.023	-.136*	.113*	.141*	-.036*	.026	-.018
		(.000)	(.002)	(.000)	(.000)	(.063)	(.001)	(.310)	(.559)	(.675)	(.091)	(.064)	(.061)	(.020)	(.310)	(.556)
	Anger	-.411*	.235*	.624*	-.244*	.167	.261*	-.131*	-.031	.049	.130*	-.081	-.069	.062*	-.082*	.008
		(.000)	(.000)	(.000)	(.000)	(.158)	(.001)	(.098)	(.826)	(.395)	(.094)	(.136)	(.156)	(.004)	(.097)	(.824)
Participant is Exploited	Joy	.496*	-.149*	.065	.267*	-.573*	.203*	-.150*	-.042	-.336*	-.001	-.335*	-.284*	-.030*	-.010	-.011
		(.000)	(.003)	(.132)	(.000)	(.000)	(.000)	(.025)	(.716)	(.000)	(.985)	(.000)	(.000)	(.022)	(.208)	(.714)
	Guilt	-.480*	.469*	-.033	-.282*	-.450*	.127*	-.125*	-.034	.521*	.232*	.289*	.216*	.059*	.004	.009
		(.000)	(.000)	(.548)	(.000)	(.000)	(.040)	(.058)	(.769)	(.000)	(.003)	(.000)	(.000)	(.044)	(.566)	(.767)
Participant Exploits	Sad	-.136*	.393*	.153*	-.186*	.624*	.016	.066	-.127	-.076	-.031	-.045	-.085*	.006	.010	.024
		(.000)	(.000)	(.004)	(.000)	(.000)	(.809)	(.299)	(.311)	(.143)	(.599)	(.223)	(.001)	(.808)	(.323)	(.311)
	Anger	-.196*	.034	.543*	-.182*	.621*	-.009	.106	-.131	-.112*	-.071	-.041	-.122*	-.000	.058	.024
		(.000)	(.500)	(.000)	(.000)	(.000)	(.879)	(.139)	(.284)	(.038)	(.272)	(.330)	(.000)	(.881)	(.138)	(.286)

Note. Cn = Conduciveness to goals; SB = Self-Blame; PB = Participant-Blame; CP = Coping Potential. Values correspond to unstandardized regression coefficients (*p* values in parentheses).

* $p < .05$.

3.3.2 Experiment 2: Establishing the Causal Model

Spencer, Zanna and Fong (2005) argue that showing mediation statistically, as proposed by Baron and Kenny (1986), is no substitute to showing mediation experimentally. As alternatives to the statistical approach, they propose: (a) the experimental-causal-chain design, where each link of the proposed causal model is shown experimentally; (b) the moderation-of-process design, where moderators are used to manipulate the mediators. In this dissertation, we focus on the first approach and leave the moderation-of-process approach as a possible topic of future work. Applying the experimental-causal-chain approach to our case means showing, experimentally, each of the causal links in the proposed causal model (Figure 3.9). The effect of emotion displays on perception of appraisals (causal link i) was already shown, experimentally, in Experiment 1. In turn, the goal of Experiment 2 was to show experimentally the effect of perception of appraisals on perceptions of cooperativeness (causal link ii). To accomplish this, perceptions of appraisals were manipulated by having the agents, instead of displaying emotion through the face, express the appraisals directly through text. Since our argument is that the information people get from emotion displays corresponds to perception of appraisals, we hypothesized that this new manipulation would lead to similar effects on perception of cooperativeness to the ones reported in Experiment 1 (H3.4).

3.3.2.1 Method

The scenarios and game remained the same as in Experiment 1. Regarding conditions, Experiment 2 also followed a mixed design with two factors: Outcome (between-participants) with 4 levels (one for each possible outcome); Appraisals (repeated-measures) with 5 levels (Neutral vs. Joy vs. Anger vs. Sadness vs. Guilt). Only the pairings of outcome and appraisals explored in Experiment 1 were used in this study. The manipulation consisted, instead of emotion

displays, of textual expression of the appraisals. According to appraisal theories (Ellsworth & Scherer, 2003), three appraisals are relevant for the explored emotions: conduciveness to goals, blameworthiness and coping potential. However, since participants in Experiment 1 did not seem to be interpreting coping potential as predicted by some appraisal theories (see the discussion for Experiment 1) and it was the least relevant mediator (Table 3.16), we focused only on conduciveness to goals and blameworthiness. The mapping of emotion into appraisals is shown in Table 3.17. Participants were still introduced to the agents they imagined playing with, however, only a static image was shown of the (neutral) face. The textual expression of appraisals was simulated by typing at the bottom of the screen, as if simulating a chat interface (Figure 3.14).

Table 3.17. Mapping of emotions to textual expression of appraisals (Experiment 2, Study 3).

Emotion	Appraisal Expression
Neutral	I neither like, nor dislike this outcome
Joy	I like this outcome
Anger	I do NOT like this outcome and I blame YOU for it
Sadness	I do NOT like this outcome
Guilt	I do NOT like this outcome and I blame MYSELF for it

Regarding measures, after watching the video of the agent's reaction, we asked participants the same questions about perception of appraisals and cooperativeness as in Experiment 1. We omitted the questions regarding emotion interpretation in this study.

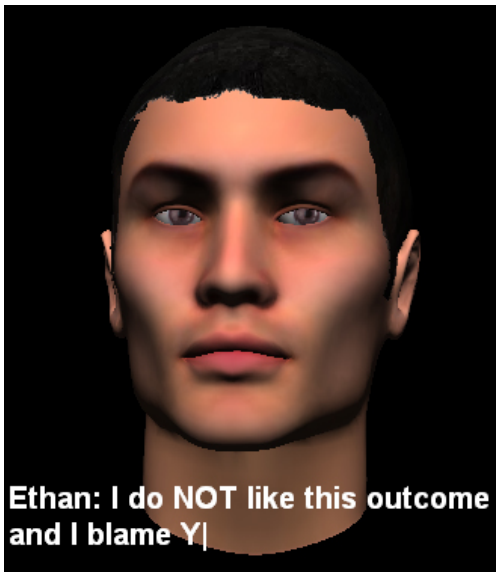


Figure 3.14. Textual expression of appraisals in Experiment 2 (Study 3).

We recruited two-hundred and two (*N*=202) participants online using Amazon Mechanical Turk. This resulted in approximately 50 participants for each outcome. Gender distribution was as follows: *males*, 51.0%; *females*, 49.0%. Age distribution was as follows: *18 to 21 years*, 10.4%; *22 to 34 years*, 56.4%; *35 to 44 years*, 12.9%; *45 to 54 years*, 12.4%; *55 to 64 years*, 5.9%; *65 years and over*, 2.0%. Most participants were from the United States (66.3%) and India (22.8%). The education level distribution was as follows (current or expected degrees): *high school*, 15.3%; *college*, 62.9%; *Masters*, 18.3%; *Ph.D. or above*, 3.5%. Education majors and profession were quite diverse. Participants were paid USD $1.02 and average participation time was 25 minutes.

3.3.2.2 Results

Manipulation check. Similarly to Experiment 1, questions 1 to 4 were highly correlated (α = .853) and, thus, were collapsed (averaged) into a single measure called *conduciveness to goals*. Questions 7 to 9 were also correlated (α = .695) and were collapsed into a single variable called

coping potential. In this study, the measures for conduciveness to goals and blameworthiness could be used to check that the manipulation was successful. For each outcome, we conducted a repeated-measures ANOVA to analyze the effect of appraisals on conduciveness to goals, self-blameworthiness (question 5), participant-blameworthiness (question 6) and coping potential. Means, standard deviations, significance levels and effect sizes are reported in Table 3.18. In summary: (a) in mutual cooperation, participants perceived the agent that expressed joy to find the outcome more conducive, blame less, and have higher coping potential than the agent that expressed no emotion; (b) in mutual defection, the agent that expressed joy was perceived as having higher conduciveness to goals than the agent that showed no emotion (pairwise comparisons were based on Bonferroni post-hoc comparisons, which are not shown in Table 3.18). In turn, the neutral agent was perceived as having higher conduciveness to goals and coping potential than the angry agent. The angry agent was also perceived as blaming the participant the most and himself the least; (c) when the participant was exploited, participants perceived the agent that expressed joy to find the outcome more conducive and have higher coping potential than the neutral agent. The neutral agent, in turn, was perceived as having higher goal conduciveness and coping potential than the agent that expressed guilt. Finally, the agent that expressed guilt was perceived as blaming himself the most; (d) when the participant exploited, the agent that expressed no emotion was perceived as having higher conduciveness to goals and coping potential than the angry or sad agents. The angry agent blamed the participant the most and the sad agent blamed himself the most. Overall, the results on perception of conduciveness to goals and blameworthiness suggested that the manipulation was successful. Notice also that the results on coping potential were very similar to the results in Experiment 1, suggesting that people were able to make inferences about coping potential given only information about the other two appraisals.

Table 3.18. Descriptive statistics for perceptions of appraisals in Experiment 2 (Study 3).

	Conduciveness to Goals	Blame (Self)	Blame (Participant)	Coping Potential
Mutual cooperation (n=52)				
Neutral	4.08 (0.805)	3.35 (1.399)	3.62 (1.611)	4.35 (0.668)
Joy	5.71 (0.779)	2.63 (1.749)	2.94 (1.809)	5.02 (0.823)
Sig. (r)	.000* (.892)	.002* (.418)	.018* (.325)	.000* (.682)
Participant is exploited (n=52)				
Neutral	4.30 (0.711)	2.69 (1.566)	2.67 (1.568)	4.74 (0.968)
Joy	5.92 (1.007)	3.37 (2.151)	2.73 (1.784)	5.50 (0.953)
Guilt	3.08 (0.977)	6.29 (1.377)	1.98 (1.407)	3.74 (1.149)
Sig. (partial η^2)	.000* (.711)	.000* (.578)	.009* (.088)	.000* (.517)
Participant exploits (n=48)				
Neutral	3.91 (1.021)	2.85 (1.750)	3.33 (1.826)	4.90 (0.899)
Anger	2.52 (1.357)	2.08 (1.648)	6.40 (1.005)	3.54 (1.195)
Sadness	2.75 (1.338)	3.85 (1.750)	4.83 (1.906)	4.17 (1.192)
Sig. (partial η^2)	.000* (.490)	.000* (.360)	.000* (.483)	.000* (.444)
Mutual defection (n=50)				
Neutral	4.20 (0.713)	3.18 (1.438)	3.18 (1.466)	4.51 (0.747)
Joy	5.42 (0.937)	2.78 (1.067)	2.72 (1.565)	4.80 (0.898)
Anger	2.26 (1.064)	1.62 (1.067)	6.44 (1.264)	2.86 (0.953)
Sig. (partial η^2)	.000* (.760)	.000* (.309)	.000* (.656)	.000* (.640)

* $p < .05$.

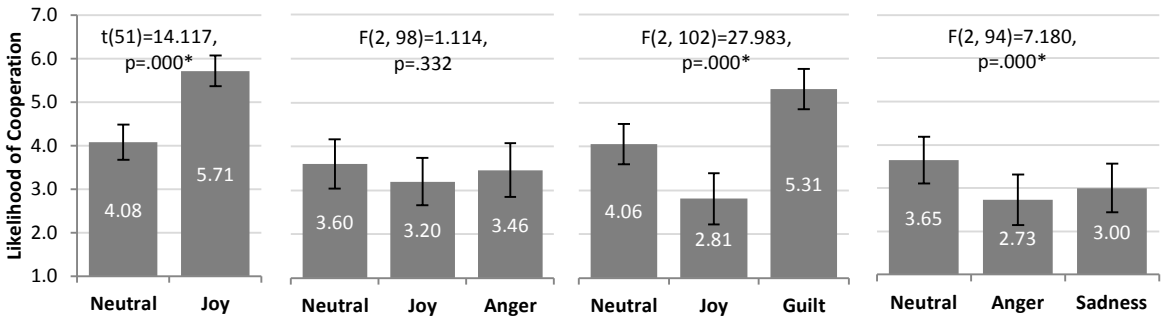
Effects for perception of cooperativeness. For each outcome, we conducted a repeated-measures ANOVA to analyze the effect of emotion display on perception of cooperativeness. Figure 3.15 shows the means and standard errors. Table 3.19 reports means, standard deviations, significance levels and effect sizes. If we collapse the data from the two experiments, it is also possible to analyze whether there was any interaction between sample and emotion displays. Because the argument is that appraisal expressions are part of the information retrieved from emotion displays, we expected there to be no interactions. Table 3.19 also shows these interactions. In summary: (a) in mutual cooperation, the agent that expressed joy was perceived as being more cooperative than the neutral agent; (b) in mutual defection, there were no effects; (c) when the

participant was exploited, the agent that expressed guilt was perceived as more cooperative than the agent that expressed no emotion (pairwise comparisons were based on Bonferroni post-hoc comparisons, which are not shown in Table 3.19). In turn, the agent that expressed no emotion was perceived as more cooperative than the agent that expressed joy; (d) when the participant exploited, the agent that expressed no emotion was perceived as more cooperative than the agent that expressed sadness or anger. Overall, the effects on perception of cooperativeness were very similar to the effects reported in Experiment 1 thus, confirming hypothesis H3.4. Moreover, as expected, there were no significant interactions between sample and emotion display.

Table 3.19. Descriptive statistics for perceptions of cooperativeness in Experiment 2 (Study 3).

Mutual cooperation (n=52)		*Mutual defection (n=50)*	
Neutral	3.27 (1.693)	Neutral	3.60 (1.990)
Joy	4.85 (1.841)	Joy	3.20 (1.874)
-	-	Anger	3.46 (2.159)
Sig. (*r*)	.000* (.654)	Sig. (η_p^2)	.332 (.022)
Sample x Emotion, Sig. (*r*)	.267 (.005)	Sample x Emotion, Sig. (*r*)	.701 (.002)
Participant is Exploited (n=52)		*Participant exploits (n=48)*	
Neutral	4.06 (1.650)	Neutral	3.65 (1.839)
Joy	2.81 (2.077)	Anger	2.73 (2.029)
Guilt	5.31 (1.639)	Sadness	3.00 (1.935)
Sig. (η_p^2)	.000* (.354)	Sig. (η_p^2)	.001* (.133)
Sample x Emotion, Sig. (*r*)	.473 (.005)	Sample x Emotion, Sig. (*r*)	.963 (.000)

* $p < .05$.



(a) Mutual Cooperation (b) Mutual Defection (c) Participant is Exploited (d) Participant Exploits

Figure 3.15. Perception of cooperativeness in Experiment 2 (Study 3).

3.3.2.3 Discussion

The results showed that people's perceptions about the agent's intentions were influenced by how people perceived the agent to be appraising the outcomes. Moreover, the experiment showed that conveying information about the appraisals corresponding to the emotions explored in Experiment 1, led to similar effects on perceptions of cooperativeness as in Experiment 1: if the agent expressed that an unfavorable outcome was obstructive, then it was perceived as being more unwilling to cooperate in the future than if it expressed a neutral message; if, in addition, the agent blamed the participant for the unfavorable outcome, then it was perceived as even less willing to cooperate than in the previous case; if the agent got a favorable outcome at the expense of the participant, communicating self-blame for the outcome signaled that the agent was regretful and willing to cooperate in the future; finally, if the agent communicated liking the outcome, then it was perceived as more or less willing to cooperate in the future according to context. Furthermore, when collapsing the data from both experiments, there were no interactions between sample and emotion display on perceptions of cooperativeness, i.e., independently of whether appraisals were conveyed through facial displays or directly by text, the effect on perception of cooperativeness was the same. This suggests that, in line with the reverse appraisal proposal, the information people get from emotion displays pertains to how the agent is

appraising the outcomes. Finally, according to Spencer et al. (2005), since Experiment 1 showed that emotion displays influence perceptions of appraisal and Experiment 2 showed that perceptions of appraisal influence perceptions of cooperativeness, there is experimental evidence in support of the causal model proposed in Figure 3.9.

3.3.3 General Discussion

This study presented two experiments that show that emotion displays influence people's perceptions about the agent's likelihood of cooperation in a social dilemma. Experiment 1 showed that: (a) emotion displays influenced how participants perceived the agent showing emotion to be appraising social dilemma outcomes; (b) emotion displays influenced participant's perception of how likely the agent showing emotion was to cooperate in the future; (c) perceptions of appraisals mediated the effect of emotion displays on perceptions of cooperativeness. Experiment 2 showed that: (a) manipulating how one perceived the agent to be appraising social dilemma outcomes influenced how participants perceived the agent to be likely to cooperate in the future; (b) if the manipulation of appraisal perceptions corresponded to the emotions in Experiment 1, the effect on perception of cooperativeness was identical. Altogether, these experiments provide evidence for a causal model (Figure 3.9) where emotion displays cause one to infer how the agent is appraising the social dilemma outcomes, which in turn causes one to infer how likely the agent is to cooperate in the future. This model is in line with the social-functions view of emotions–that argues emotions communicate information about one's beliefs, desires and intentions–and, in particular, with the reverse appraisal proposal–that argues people infer, from emotion displays, how others are appraising the situation, which in turns, supports inferences about others' mental states.

Chapter Four: Computational Models

Having established the theoretical foundations for the social effects of emotion on people's decision-making, this chapter describes a series of computational models for decision-making in the prisoner's dilemma. These models were developed using statistical and machine learning techniques on the data collected in the empirical studies. With these models we sought to: (a) gain insight into the effects reported in the previous chapter; (b) establish further theoretical results about the relevance of emotion and appraisals for computational models of decision-making; (c) develop models that, pragmatically, could be used to drive an agent which was engaged in the prisoner's dilemma (with a human or another agent).

4.1 Modeling the Effect of Emotion Displays

4.1.1 Overview

In this section we describe a computer model of decision-making (de Melo, Carnevale, Antos, & Gratch, 2011) that takes into account the counterpart's emotion displays and replicates findings from the first two studies reported in the previous chapter. Methodologically, we followed a data-driven approach: (1) Data from the studies were collapsed into a single database. Features represented what happened in the round and whether the participant cooperated in the next round (target); (2) Probabilistic models were fitted to the data using maximum likelihood estimation (Alpaydin, 2010). Each model predicted likelihood of cooperation given a subset of the features (e.g., outcome and display in the current round). We explored models that predicted based on outcome only, outcome and emotion, and outcome, emotion and contrasts (i.e., order of play); (3) Regarding evaluation, even though we looked at standard performance measures such as error rate, the focus was on the models' ability to replicate how people behaved in the prisoner's

dilemma. In order to accomplish this, we "played" the models with different agents that displayed emotions under the same configurations as in the original studies. Given the findings reported in the previous chapter, our hypotheses were that: a model that considered emotion displays would outperform a model that did not (H4.1); a model that considered emotion and order of play would outperform a model that only considered emotion (H4.2).

4.1.2 Data and Features

The data consisted of examples corresponding to each round each participant played in Studies 1 and 2 (see Sections 3.1 and 3.2 for a detailed description of these studies). Data corresponding to last rounds was ignored, since the goal was to predict whether the participant would cooperate in the next round. In total, there were 12,432 examples[4]. The feature set was the following:

(a) Outcome of the Round: whether the players cooperated or defected;

(b) Emotion Display: the agent's display following the outcome in that round;

(c) First Game: whether the example corresponded to a round in the first game;

(d) Agent is cooperator: 'true' if current agent was a cooperator;

(e) Previous Agent is Cooperator: 'true' if (eventual) previous agent was a cooperator;

(f) Whether Participant Cooperates in the Next Round: this is the target attribute.

4.1.3 Training, Validation and Test Sets

The data was first partitioned into a training (75%) and a test set (25%). The training set was further partitioned into 20 subsets to support 20-fold cross-validation. Every subset (including the test set) was created while making sure they had the same proportion of positive and negative examples from each of the empirical studies.

[4] Data from a third study, which led to similar findings as Study 2 and is not described in the dissertation, was also used. Further details can be found in de Melo, Carnevale, Antos, et al. (2011).

4.1.4 Models

Models consisted of rules defining the probability of cooperation in the next round, given a subset of the features. We explored three different model variants, described below, that use different subsets of the features. Maximum likelihood estimation was used to fit the models to the data and estimate the parameters (Alpaydin, 2010). The training procedure consisted of 20-fold cross-validation and the final model parameters corresponded to the average over all training sets.

Model 1: Model based on Outcome. The first model predicted likelihood of cooperation based only on outcome of the current round. Outcome was chosen as the first attribute as it ranked best according to the information gain metric (or Kullback-Leibler divergence; Alpaydin, 2010). Thus, the model predicted probability of cooperation in the next round given a certain outcome in the present round. These probabilities were obtained by calculating the frequency the participant cooperated after each round, for each outcome. Table 4.1 shows the parameters (averaged over all training sets) for this model (under Model 1).

Model 2: Model based on Outcome and Emotion. The next model predicted likelihood of cooperation given outcome and the agent's display. This model's parameters are shown in Table 4.1 (Model 2).

Model 3: Model based on Outcome, Emotion and Order of Play. Finally, the third model also predicted likelihood of cooperation based on outcome and emotion displays but, also took into account the black-hat/white-hat contrast effects reported in the empirical studies. All the information required to represent these effects was in attributes (c), (d) and (e), i.e., attributes regarding whether the first and second agents were black-hats (non-cooperators) or white-hat

(cooperators). Notice, however, these attributes are conceptually different than outcome and emotion displays, because they were *non-observable*. Effectively, they represent *inferences* participants made while playing the games. Nevertheless, notice these inferences *were* made, consciously or not, because otherwise there would have been no contrast effects. Still, modeling the mechanism by which participants make these inferences was left for future work and, for the time being, we simply assumed that attributes (c), (d) and (e) were directly observable. In summary, the third model calculated, for each combination of the attributes (c), (d) and (e), probabilities given the outcome and agent's displays in the previous round (see Table 4.1 under Model 3).

Table 4.1. Parameters for the maximum likelihood models.

				Model 3				
	Emotion	Model 1	Model 2	BH *1st Game*	WH *1st Game*	BH→WH *2nd Game*	WH→BH *2nd Game*	BH→BH *2nd Game*
CC	joy	.67	.72		.64	.76		
	neutral		.62	.53			.71	.51
	sadness		.61	.61			.57	.54
DD	neutral	.22	.24	.22	.27	.30	.21	.26
	sadness		.20	.21	.20	.20	.17	.25
C_HD_A	joy	.29	.26	.30			.27	.16
	neutral		.26	.26	.15	.40	.19	.23
	sadness		.34		.35	.33		
	guilt		.36		.27	.40		
D_HC_A	anger	.28	.27	.27	.28	.27	.37	.23
	neutral		.24	.22			.34	.19
	sadness		.31	.27	.30	.29	.34	.37

Note. CC = mutual cooperation; DD = mutual defection; C_HD_A = human cooperates, agent defects; D_HC_A = human defects, agent cooperates; BH = Black-Hat (or non-cooperator); WH = White-Hat (or cooperator); *1st Game* refers to probabilities in the 1st game (with a BH or WH); *2nd Game* refers to probabilities in the second game (with a BH or WH) but, when the game was preceded by a specific first game (with another BH or WH). Notice there is no prediction for the case where both the 1st and 2nd agents are white-hats because this was not explored in our studies. Values in the table represent probabilities of cooperation.

4.1.5 Model Selection

Model selection was based on minimizing error rate, i.e., the percentage of incorrectly classified examples (averaged over all 20 validation sets). Table 4.2 shows the error rates for each model. The results showed that error rates were significantly different ($F(2, 57) = 28.207$, $p < .05$) and, LSD post-hoc tests revealed that: the error rate for Model 1 was higher than for Model 2 ($p = .100$); and, the error rate for Model 2 was higher than for Model 3 ($p = .000$). Table 4.2 also reports several other standard measures–precision, recall, F1, and so forth (Alpaydin, 2010)–and it is clear that Model 3 outperformed Model 2 which, in turn, outperformed Model 1. Table 4.3 reports the results over the test set. Error rate suggests, once again, that Model 3 performed better than Model 2 and, in turn, Model 2 performed better than Model 1. The remaining variables in Table 4.3 also generally support that Model 3 fared best and that Model 1 fared worst. Finally, *average log likelihood* measures the posterior probability of the (whole) dataset given the model, averaged over the number of examples (the closer to 0, the better). The results for the models were: Model 1, *-0.247*; Model 2, *-0.246*; and, Model 3, *-0.245*. Thus, the results suggested that the data was most likely to have been generated from Model 3 than any of the other models.

Table 4.2. Performance measures over validation sets.

	Model 1		Model 2		Model 3			
	Mean	SD	Mean	SD	Mean	SD	F	Sig.
error	.382	.016	.373	.017	.345	.017	28.207	.000*
accuracy	.618	.016	.627	.017	.655	.017	28.207	.000*
precision	.408	.024	.422	.025	.466	.025	29.842	.000*
recall	.407	.025	.423	.026	.466	.024	29.571	.000*
F1	.408	.025	.423	.026	.466	.024	29.575	.000*
true positives	61.332	4.935	63.717	5.298	70.134	4.958	16.147	.000*
false positives	88.885	4.434	87.196	4.582	80.339	4.785	19.342	.000*
true negatives	226.566	3.599	228.254	3.760	235.112	3.886	29.136	.000*
false negatives	89.069	4.618	86.684	4.526	80.267	4.583	19.798	.000*

* $p < .05$.

Table 4.3. Performance measures over the test set.

model	error	accuracy	precision	recall	F1	tp	fp	tn	fn
Model 1	.382	.618	.411	.421	.416	422.86	606.99	1498.01	581.14
Model 2	.380	.620	.414	.425	.419	426.72	605.25	1499.75	577.28
Model 3	.378	.622	.417	.424	.421	425.85	595.55	1509.45	578.15

Note. tp = true positives; fp = false positives; tn = true negatives; fn = false negatives.

4.1.6 Evaluation

The results in the previous subsection suggested Model 3 was best and Model 1 worst at predicting how humans behaved in the prisoner's dilemma. However, in this subsection we explicitly test this by replicating Studies 1 and 2 but, substituting humans for the maximum likelihood models. Aside from verifying the results from the previous subsection, this experiment allowed us to get insight into the mechanisms that explain why some models fared better than others. To accomplish this, we ran each model 1000 times (500 times per order) for each experiment in our studies, and measured which findings were replicated by the model. The cooperation rates and standard deviations for the original human data and the models are shown in Table 4.1. Two columns are shaded in this table, for each model: (1) the left column summarizes whether cooperation rates were significantly different ($p < .05$) and represented an effect size above a minimum threshold[5], which we set to 1.5 (corresponding to, at least, a small effect size). For instance, a '>' means the model cooperated significantly more with the agent on the left than the agent on the right and the effect size passes the threshold; (2) the right column shows a tick if the model successfully replicated the findings in the human data. Therefore, the more ticks a model has, the better it was at replicating findings. Overall, the percentage of findings each model replicated was: Model 1, *50.0%* (6 out of 12 ticks); Model 2, *75.0%* (9 out of 12 ticks); and, Model 3, *100.0%* (12 out of 12 ticks).

[5] Because it is possible to get significance even for small differences if the sample size is large enough, it is important to require the effect size to be above a minimum threshold.

4.1.7 Discussion

In this section we proposed a data-driven probabilistic model for decision-making in the prisoner's dilemma that took into account the counterpart's emotion displays. The evaluation revealed that the model was better at replicating findings about how humans behave in the prisoner's dilemma if, instead of considering round outcome alone, it also considered emotion displays. Thus, hypothesis H4.1 was confirmed. This is also in line with the findings reported in the previous chapter regarding the importance of emotion displays. Furthermore, the results confirmed that considering (black-hat/white-hat) contrast effects further improved the model's ability to predict human behavior; therefore, hypothesis H4.2 was also verified. Theoretically, the model complements the findings in the empirical studies by quantizing (through probabilities) the effect of emotion displays on decision-making in the prisoner's dilemma. For instance, Model 2 (see Table 4.1) suggests that, after the human was exploited by the agent (i.e., when the human cooperated and the agent defected), the human's likelihood of cooperating goes up from 26% to 36% if the agent displayed guilt as opposed to joy. Finally, pragmatically, the model can be used to drive an agent that is engaged in the prisoner's dilemma with another party (human or agent) that shows emotion.

Table 4.4. Evaluation of the maximum likelihood models.

		Humans			**Model 1** Outcome				**Model 2** Outcome & Emotion				**Model 3** Outcome & Emotion & Order of Play			
Study 1	*Order*	*Cooperative*	*Individual.*		*Cooperative*	*Individual.*			*Cooperative*	*Individual.*			*Cooperative*	*Individual.*		
Coop	both	.37 (.28)	.27 (.23)	>	.33 (.14)	.33 (.14)	≈	×	.36 (.16)	.31 (.13)	>	✓	.35 (.16)	.31 (.14)	>	✓
Vs.	coop→comp	.35 (.26)	.31 (.26)	≈	.32 (.14)	.33 (.14)	≈	✓	.37 (.17)	.30 (.12)	>	×	.31 (.14)	.32 (.15)	≈	✓
Comp	comp→coop	.39 (.30)	.23 (.19)	>	.33 (.14)	.33 (.14)	≈	×	.35 (.16)	.31 (.13)	>	✓	.39 (.17)	.30 (.12)	>	✓
Study 2		*Cooperative*	*Competitive*		*Cooperative*	*Competitive*			*Cooperative*	*Competitive*			*Cooperative*	*Competitive*		
Coop	both	.41 (.23)	.39 (.21)	≈	.34 (.15)	.34 (.15)	≈	✓	.36 (.15)	.33 (.13)	>	✓	.39 (.16)	.35 (.15)	>	✓
Vs.	coop→comp	.37 (.18)	.49 (.19)	<	.35 (.15)	.33 (.14)	≈	×	.36 (.15)	.32 (.14)	>	×	.33 (.13)	.38 (.16)	<	✓
Ctrl	comp→coop	.44 (.25)	.32 (.20)	>	.34 (.14)	.34 (.15)	≈	×	.37 (.16)	.33 (.13)	>	✓	.46 (.17)	.31 (.12)	>	✓
		Cooperative	*Control*		*Cooperative*	*Control*			*Cooperative*	*Control*			*Cooperative*	*Control*		
Coop	both	.34 (.17)	.24 (.14)	>	.34 (.14)	.34 (.15)	≈	×	.36 (.14)	.31 (.13)	>	✓	.38 (.16)	.31 (.14)	>	✓
Vs.	coop→ctrl	.24 (.09)	.21 (.12)	≈	.34 (.14)	.34 (.14)	≈	✓	.35 (.14)	.32 (.13)	>	×	.32 (.12)	.34 (.16)	≈	✓
Ctrl	ctrl→coop	.39 (.19)	.26 (.15)	>	.33 (.14)	.33 (.15)	≈	×	.36 (.14)	.31 (.13)	>	✓	.44 (.17)	.29 (.12)	>	✓
		Competitive	*Control*		*Competitive*	*Control*			*Competitive*	*Control*			*Competitive*	*Control*		
Comp	both	.23 (.11)	.23 (.17)	≈	.35 (.15)	.34 (.14)	≈	✓	.33 (.13)	.31 (.13)	≈	✓	.29 (.11)	.29 (.11)	≈	✓
Vs.	comp→ctrl	.22 (.10)	.25 (.18)	≈	.35 (.15)	.35 (.14)	≈	✓	.33 (.12)	.31 (.12)	≈	✓	.30 (.11)	.29 (.10)	≈	✓
Ctrl	ctrl→comp	.25 (.13)	.20 (.16)	≈	.35 (.14)	.34 (.14)	≈	✓	.32 (.13)	.32 (.13)	≈	✓	.27 (.10)	.29 (.11)	≈	✓

Note. Cooperation rates (standard deviations) are shown for the original empirical data (under 'Humans') and when running the models under each of the experimental configurations. The left-most shaded column summarizes the comparison between cooperation rates between the two agents in that configuration. The right-most shaded column is interpreted as follows: ✓ means the model replicates the findings in the human data; × means the model doesn't replicate the human data.

4.2 Modeling the Effect of Appraisals

4.2.1 Overview

In this section, using data from Study 3–which, contrary to Studies 1 and 2, measured participants' perceptions of appraisal–we present a Bayesian model for decision-making in the prisoner's dilemma that shows the value of taking appraisals into account (de Melo, Carnevale, Stephen, Antos, & Gratch, 2012). At its core, the new model needs to *infer*, from emotion displays, how the counterpart appraises the situation and, from this, *infer* the other's intentions in the social encounter. Because there is a strong inductive component to the model, we followed a Bayesian approach (Griffiths, Kemp, & Tenenbaum, 2008). We considered three alternative Bayesian networks: the first considered the outcome of the dilemma only; the second considered the outcome and the emotion display; the third considered the outcome, emotion display and appraisals. We compared models with respect to their accuracy in predicting the counterpart's likelihood of cooperation in the future. Our first hypothesis, following findings in the previous section, was that: Models that considered emotion displays would have better accuracy than models that did not (H5.1).

However, the focus in this section is on showing the value of integrating (perceptions of) appraisals in a model of decision-making. One important advantage appraisals provide is a structure which is shared by several emotions. For instance, conduciveness to goals is an appraisal which is shared by joy and sadness (Ellsworth & Scherer, 2003): an event which is conducive to someone's goals causes joy; an event which is obstructive to someone's goals causes sadness. This shared structure provides a mechanism for learning parameters and making inferences regarding emotions even in the absence of examples for that particular emotion. All that is necessary is data for the emotions with which the missing emotion shares appraisals. So,

our next hypothesis was: Models that considered appraisals would have better accuracy than models that did not, over test sets which included emotions which were not seen in the training set (H5.2).

Finally, there are situations where people express how they are appraising a situation without resorting to emotion expression. An obvious example is when people convey verbally their attitudes toward an event. The data collected in the second experiment in Study 3–where people convey appraisals through text–is a case in point. This dataset could, thus, be used to test our third and final hypothesis: Models that considered appraisals could be accurate even when no emotion was shown (H5.3).

4.2.2 Data and Features

The models presented here use data from both experiments in Study 3 (Section 3.3). Recall that in this study participants imagined playing the first round of the iterated prisoner's dilemma with virtual agents. In Experiment 1, the agents displayed emotions in the face; in Experiment 2, the agents expressed textually how they were appraising the ongoing interaction. In each scenario, after watching the agent's reaction, participants were asked several questions, on a 1 to 7 scale, about how was the agent appraising the interaction and how likely was the agent to cooperate in the future. For the purposes of learning a Bayesian model, the appraisal and likelihood of cooperation questions were converted into binary format: the feature was set to 'true' if the original classification was 5 or above; the feature was set to 'false' if the classification was 3 or below; if the classification was 4, the feature was not assigned a value (missing attribute). Each example in the training datasets, thus, had the following features:

a) Outcome of the Round: whether the players cooperated or defected;
b) Emotion Display: Neutral, Joy, Anger, Guilt or Sadness (dataset for Experiment 1 only);
c) Conduciveness to Goals (binary): Whether the agent was perceived to find the outcome conducive to its goals;

d) Self-Blame (binary): Whether the agent was perceived to blame itself for the outcome;

e) Participant-Blame (binary): Whether the agent was perceived to blame the participant for the outcome;

f) Coping Potential (binary): Whether the agent was perceived to be able to deal with the consequences of the outcome;

g) Likelihood of Cooperation (binary): Whether the agent was perceived to be likely to cooperate in the future.

The dataset corresponding to Experiment 1 had, excluding the examples for which the target attribute (Likelihood of Cooperation) was missing, 940 examples. The dataset corresponding to Experiment 2 had 454 examples. The main difference between Dataset 1 and Dataset 2 is that the latter did not have a feature for emotion displays (or, equivalently, its values were missing).

4.2.3 Bayesian Models

All Bayesian models were trained with respect to Dataset 1 (Experiment 1 in Study 3). Since some of the attributes in the examples were missing, the EM algorithm was used for learning the parameters (Alpaydin, 2010). The decision regarding Likelihood of Cooperation was made as follows:

- If P(Likelihood of Cooperation) > 0.5, true
- If P(Likelihood of Cooperation) = 0.5, random
- Otherwise, false

Model 1: Outcome. The first Bayesian model considered only two variables: Outcome (O) and Likelihood of Cooperation (LC). Figure 4.1 shows the respective Bayesian network. Outcome was set to have a uniform prior, i.e., each possible outcome occurred with .25 probability. The learnt parameters are shown in Table 4.5.

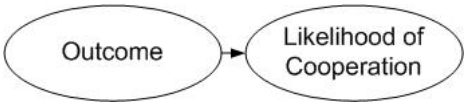


Figure 4.1. Bayesian network for Model 1.

Table 4.5. Parameters for Bayesian Model 1.

O	P(LC)	O	P(LC)
CC	.470	C_HD_A	.380
DD	.405	D_HC_A	.271

Model 2: Emotion and Outcome. The second Bayesian model built on the previous and added Emotion Display (ED). Figure 4.2 shows the respective Bayesian network. Emotion Display was also set to have a uniform prior, i.e., each emotion occurred with .20 probability. The parameters are shown in Table 4.6.

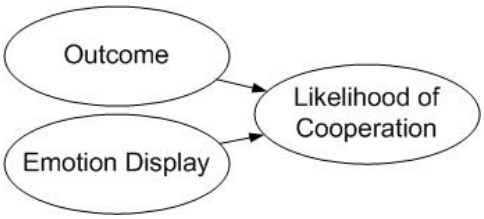


Figure 4.2. Bayesian network for Model 2.

Table 4.6. Parameters for Bayesian Model 2.

ED	**O**	**P(LC)**	**O**	**P(LC)**
Neutral	CC	.235	C_HD_A	.254
Joy	CC	.719	C_HD_A	.182
Anger	CC	.500	C_HD_A	.500
Guilt	CC	.500	C_HD_A	.670
Sadness	CC	.500	C_HD_A	.500
Neutral	DD	.453	D_HC_A	.377
Joy	DD	.368	D_HC_A	.500
Anger	DD	.400	D_HC_A	.242
Guilt	DD	.500	D_HC_A	.500
Sadness	DD	.500	D_HC_A	.217

Model 3: Appraisals. The last Bayesian model added appraisal variables: Conduciveness to Goals (CG), Self-Blame (SB), Participant-Blame (PB) and Coping Potential (CP). The respective Bayesian network is shown in Figure 4.3. The appraisal variables were given BDeu priors (Heckerman, Geiger, & Chickering, 1995), i.e., likelihood equivalent uniform Dirichlet priors. The parameters for the appraisal variables are shown in Table 4.7 and the parameters for Likelihood of Cooperation in Table 4.8.

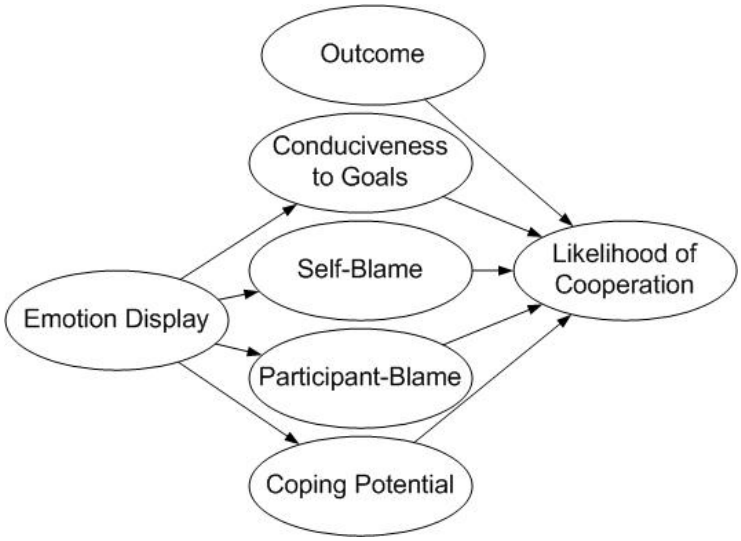


Figure 4.3. Bayesian network for Model 3.

Table 4.7. Parameters for the appraisal variables in Bayesian Model 3.

ED	P(CG)	P(SB)	P(PB)	P(CP)
Neutral	.370	.203	.267	.748
Joy	.970	.206	.177	.905
Anger	.021	.381	.824	.324
Guilt	.227	.678	.222	.348
Sadness	.041	.730	.485	.285

Table 4.8. Parameters for Likelihood of Cooperation in Bayesian Model 3.

CG	SB	PB	CP	O	P(LC)	O	P(LC)
T	T	T	T	CC	.436	DD	.367
F	T	T	T	CC	.082	DD	.476
T	F	T	T	CC	.410	DD	.459
F	F	T	T	CC	.129	DD	.265
T	T	F	T	CC	.837	DD	.263
F	T	F	T	CC	.002	DD	.658
T	F	F	T	CC	.640	DD	.387
F	F	F	T	CC	.324	DD	.369
T	T	T	F	CC	.146	DD	.080
F	T	T	F	CC	.259	DD	.670
T	F	T	F	CC	.054	DD	.018
F	F	T	F	CC	.172	DD	.307
T	T	F	F	CC	.990	DD	.971
F	T	F	F	CC	.014	DD	.371
T	F	F	F	CC	.776	DD	.635
F	F	F	F	CC	.203	DD	.367
T	T	T	T	C_HD_A	.320	D_HC_A	.913
F	T	T	T	C_HD_A	.849	D_HC_A	.411
T	F	T	T	C_HD_A	.084	D_HC_A	.386
F	F	T	T	C_HD_A	.528	D_HC_A	.150
T	T	F	T	C_HD_A	.108	D_HC_A	.602
F	T	F	T	C_HD_A	.863	D_HC_A	.156
T	F	F	T	C_HD_A	.243	D_HC_A	.464
F	F	F	T	C_HD_A	.526	D_HC_A	.338
T	T	T	F	C_HD_A	.502	D_HC_A	.012
F	T	T	F	C_HD_A	.366	D_HC_A	.275
T	F	T	F	C_HD_A	.335	D_HC_A	.201
F	F	T	F	C_HD_A	.383	D_HC_A	.212
T	T	F	F	C_HD_A	.642	D_HC_A	.982
F	T	F	F	C_HD_A	.821	D_HC_A	.185
T	F	F	F	C_HD_A	.122	D_HC_A	.926
F	F	F	F	C_HD_A	.398	D_HC_A	.149

4.2.4 Evaluation

4.2.4.1 Experiment 1

To test hypothesis H5.1, that models which considered emotion would do better than models that did not, we tested the models accuracy with respect to Dataset 1 (data from Experiment 1 in Study 3). Each model was re-trained using 20-fold cross-validation. The models were then compared with respect to average performance on the 20 test sets. Several standard performance measures are reported in Table 4.9. Means were compared using the one-way independent ANOVA test.

Table 4.9. Bayesian models performance results in Experiment 1.

	acc	**tp**	**tn**	**fp**	**fn**
Model 1: Outcome	62.38%	0.00	28.75	0.00	17.25
	(5.84)	(0.00)	(3.05)	(0.00)	(2.43)
Model 2: Outcome & Emotion	69.91%	6.15	26.05	2.70	11.10
	(7.19)	(2.08)	(3.51)	(1.66)	(3.16)
Model 3: Outcome, Emotion & Appraisals	69.91%	6.15	26.05	2.70	11.10
	(7.19)	(2.08)	(3.51)	(1.66)	(3.16)
Sig. (2-sd)	*.001**	*.000**	*.013**	*.000**	*.000**

Note. acc = accuracy; tp = true positives; tn = true negatives; fp = false positives; fn = false negatives
Standard deviations are shown in parenthesis.
* $p < .05$.

The results showed that there was a significant difference in accuracy. In order to perform pairwise comparisons between the models, LSD post-hoc tests were performed (these are not shown in Table 4.9). The tests indicated that Models 2 and 3 were more accurate than Model 1. This confirmed hypothesis H5.1. Moreover, looking at the table, it is clear that Model 1 (based on Outcome) was making the same predictions as a game-theoretic model which always predicts defection (see Section 2.1). Therefore, Outcome, by itself, was insufficient to discriminate

examples in this dataset. Finally, Models 2 and 3 were also identical in their predictions. This suggests that, in this case, appraisal variables did not add more information than that provided by Emotion Display. The results also showed significant differences in the remaining variables. Looking at the true and false positive measures, it is confirmed that Model 1 always predicted defection.

4.2.4.2 Experiment 2

To test hypothesis H5.2, that the appraisal model would have better accuracy than the others over a test set with unseen emotions, we split the data in Dataset 1 into two subsets: (a) the *training subset*, which included all the examples from Experiment 1 except the ones corresponding to Joy with the outcome $C_H D_A$; (b) the *test subset*, which included all the examples from Experiment 1 where the emotion was Joy and the outcome was $C_H D_A$. Models were then trained on the former and tested on the latter. The results are shown in Table 4.10.

Table 4.10. Bayesian models performance results in Experiment 2.

	acc	**tp**	**tn**	**fp**	**fn**
Model 1: Outcome	81.82%	0.00	72.00	0.00	16.00
Model 2: Outcome, Emotion	56.82%	9.00	41.00	31.00	7.00
Model 3: Outcome, Emotion & Appraisals	81.82%	0.00	72.00	0.00	16.00

Note. acc = accuracy; tp = true positives; tn = true negatives; fp = false positives; fn = false negatives.

The results showed that Model 3 performed better than Model 2. This happened because, since there were no examples in the training set corresponding to Joy in $C_H D_A$, Model 2's posterior for Likelihood of Cooperation was .500, which corresponded to a random decision. On the other hand, because of the shared appraisal structure, Model 3's posterior for Likelihood of Cooperation ($P(LC|Joy, C_H D_A)$) was .272. Therefore, the posterior was reflecting other examples which had information about the appraisals underlying Joy. Thus, hypothesis H5.2 was

confirmed. Finally, the results revealed that, in this case, Model 1 performed as well as Model 3. This happened because both always defected in this test set.

4.2.4.3 Experiment 3

To test hypothesis H5.3, that the appraisal model could make accurate predictions even in the absence of evidence for emotion displays, we tested our models with Dataset 2 (data from Experiment 2 in Study 3). The models were still trained on Dataset 1 but, were tested on Dataset 2. The results are shown in Table 4.11.

Table 4.11. Bayesian models performance results in Experiment 3.

	acc	**tp**	**tn**	**fp**	**fn**
Model 1: Outcome	57.49%	0.00	261.00	0.00	193.00
Model 2: Outcome & Emotion	57.49%	0.00	261.00	0.00	193.00
Model 3: Outcome, Emotion & Appraisals	66.74%	72.00	231.00	30.00	121.00

Note. acc = accuracy; tp = true positives; tn = true negatives; fp = false positives; fn = false negatives.

The results showed that Model 3 outperformed the remaining models on this dataset, which confirmed hypothesis H5.3. Effectively, in the absence of information about emotion displays, Model 2 could not do better than advance a prediction based only on Outcome as in Model 1.

4.2.5 Discussion

This section confirmed that appraisals constitute a useful framework for a computational model of emotion interpretation. Following empirical results that suggest that appraisals mediated the effect of emotion displays in decision-making (Study 3), the proposed Bayesian model was structured so that variables which represented inferences about the counterpart's intentions were conditionally independent of emotion displays given information about the appraisal variables. The underlying assumption was that what mattered was not the emotion display in itself but, the

information it conveyed about appraisals. There were several advantages in developing a model based on appraisals. First, appraisals provided a structure which was shared by several emotions. This constituted a mechanism for learning parameters and making inferences regarding emotions even in the absence of examples for that particular emotion. The results in the evaluation's Experiment 2 showed that the appraisal model was capable of recovering a reasonable posterior for Likelihood of Cooperation given Joy and $C_H D_A$, even when no examples for that case existed in the training set. On the other hand, the model based on emotion and outcome (Model 2) could not do better than predict an even chance (.500) of cooperation for the case where Joy was shown in $C_H D_A$. A second advantage was that the appraisals model was capable of supporting inferences about the counterpart's intentions even in the absence of emotion. The results shown in the evaluation's Experiment 3 show that this model was capable of accurately predicting Likelihood of Cooperation for a dataset where Emotion Display was unobservable and only evidence for appraisals was available. Finally, from a cognitive modeling perspective, it is also interesting to note that the parameters for the appraisal variables (Table 4.7), which represent the conditional probabilities given the emotion display, were generally in line with expectations from appraisal theories (Ellsworth & Scherer, 2003; also see Section 2.5): conduciveness to goals was highest for joy (*P(CG|Joy)=.970*); self-blame was highest for guilt (*P(SB|Guilt)=.678*) and sadness (*P(SB|Sadness)=.730*); participant-blame was highest for anger (*P(PB|Anger)=.824*); and, coping potential was highest for Joy (*P(CP|Joy)=.905*). This means the model was capable of learning, from empirical data, some of the theoretical predictions advanced by appraisal researchers (Ellsworth & Scherer, 2003).

Chapter Five: Discussion and Implications

5.1 Summary and Contributions

To sum up, the dissertation contributes the following:

- *Empirical evidence that people's decision to cooperate in social dilemmas is influenced by emotion displayed by computer agents.* Several empirical studies were presented were participants played the finite iterated prisoner's dilemma with computer agents, that even though following the same strategy to choose their actions, displayed different emotions, through their faces, according to the outcome of each round. The results indicate that people's decision to cooperate is, in fact, influenced by emotion displays. For instance, in Studies 1 and 2, people cooperated more with an agent which displays reflected mutual cooperation (e.g., smile when both players cooperated) than one which displays reflected selfishness (e.g., smile when it defected and the participant cooperated);
- *Empirical evidence that people infer the computer agent's beliefs, desires and intentions from its emotion displays; in particular, evidence that these inferences are accomplished through reverse appraisal*. Study 2 showed that people cooperated differently with agents that differed only in the context in which a smile was shown. This is in line with appraisal theories which argue that emotion interpretation requires understanding of the circumstances, or context, that led to the generation of the emotion. Furthermore, Study 3 showed that appraisal variables mediated the effect of emotion displays on beliefs about the agent's likelihood of cooperation in the future. Additionally, Study 3 showed that manipulating experimentally how people perceive the agent to be appraising the ongoing interaction affects how people perceive the agent to be likely to cooperate. This suggests a causal model for the

interpersonal effect of emotion where people infer from emotion displays how the agent is appraising the ongoing interaction which, in turn, leads to inferences that are relevant for decision-making, such as the agent's likelihood of cooperation;

- *Computer models of decision-making in the prisoner's dilemma that take into account the counterpart's emotion displays*. Statistical and machine learning techniques were used to develop these models based on data collected in the studies. The results showed that models improved–i.e., they better replicated human behavior–if, in addition to considering information about the outcome, they also considered the emotion displayed by the counterpart. The results also showed that considering emotion and perceptions about how the counterpart is appraising the interaction, was better than considering emotion displays alone;
- *A novel paradigm for the investigation of decision-making in human-agent and human-human interaction*. The dissertation explored the interpersonal effect of emotion using a framework where participants engaged in experimental games with virtual humans, i.e., embodied agents that are capable of expressing emotion through their bodies. The results in the empirical studies were in line with findings in the behavioral sciences thus, emphasizing the viability of this research method for basic investigation in human-human interaction. Moreover, the results extend the current state-of-the-art on the interpersonal effect of emotion in ways that are consistent with existent theory on the social functions of emotion and appraisal theories. This paradigm could easily be generalized to validate and extend other theories in human-human and human-agent interaction.

5.2 Implications for Human-Computer Interaction

The dissertation presents clear evidence that it is not the mere presence of emotion but the context and information conveyed by emotion that has the potential to enhance human-computer interaction. This evidence is, thus, in contradiction with the prevalent view, we refer to as the

affective persona effect, which suggests that the mere presence of agents that express emotion enhances human-computer interaction. We argue that a social-functions view of emotion (Frijda & Mesquita, 1994; Keltner & Haidt, 1999; Keltner & Kring, 1998; Morris & Keltner, 2000; Oatley & Jenkins, 1996) is more likely to explain the interpersonal effect of emotion in human-computer interaction systems. In particular, our reverse appraisal proposal suggests that people infer the agent's beliefs, desires and intentions from its emotion displays by inferring how the agent is appraising the ongoing interaction. Conversely, if there were the possibility of recognizing the emotion being expressed by the user, computer systems could also use reverse appraisal theory to infer the user's beliefs, desires and intentions. Additionally, in line with the view that computers are social actors (Nass et al., 1994; Reeves & Nass, 1996) and the social influence model (Blascovich, 2002; Blascovich et al., 2002), the dissertation advances further empirical evidence that people can treat embodied agents like other people and be socially influenced by them. Effectively, many of the findings regarding human-agent interaction in the empirical studies were in line with previous findings regarding human-human interaction in the behavioral sciences. Finally, the dissertation advances concrete suggestions (see, for instance, Table 4.1) about which emotions an agent should express in order to promote cooperation with a human user, at least when engaged in a social dilemma.

5.3 Implications for Artificial Intelligence

Despite recent interest in emotion in artificial intelligence (Marsella et al., 2010), researchers tended to focus only on the intrapersonal effect of emotion in computer systems of decision-making. In contrast, this dissertation advances empirical evidence and a theory for the interpersonal effect of emotion in decision-making. The empirical evidence emphasizes that, at least in human-agent interaction, researchers and system designers cannot afford to neglect the *social* aspects of the interaction and, in particular, the display of emotion. At the theoretical level,

reverse appraisal is advanced as a mechanism by which people infer, from emotion displays, the other party's beliefs, desires and intentions. Integration of appraisal theories with a belief-desire-intention (BDI) architecture has already been shown to be a viable way to synthesize emotions and simulate the intrapersonal effect of emotion in decision-making in a domain-independent manner (Gratch & Marsella, 2004). Reverse appraisal complements this approach by serving as a mechanism through which agents infer other agent's beliefs, desires and intentions from their emotion displays in a domain-independent manner. Moreover, reverse appraisal emphasizes that what is critical for the social effects of emotion is not the emotion display per se but the information conveyed by the display. This is the abstract function Herbert Simon (1969) argues we should learn from biological systems and replicate in artificial systems. This is also the key to extending the current results to computational systems that go beyond virtual agents and, aside from supporting human-computer interaction, focus on agent-agent interaction.

5.4 Implications for Decision Theory

The dissertation presents evidence that departures from game-theoretic predictions commonly exhibited by people in social dilemmas are caused, at least in part, by emotional signals. The studies further emphasize the importance of context for the impact of emotion displays on decision-making. Study 2 showed that, depending on the social dilemma outcome, the same expression of a smile led to different cooperation rates. This result contrasts with research that suggests that genuine smiles are an unequivocal signal of cooperation (Brown et al., 2003; Mehu et al., 2007; Scharlemann et al., 2001) but, are in line with research which argues that a smile can have different interpretations according to context (Hareli & Hess, 2010; Van Kleef et al., 2010) and can also be shown in non-cooperative contexts (Matsumoto et al., 1986). Study 3 also showed that negative emotions can influence whether someone is perceived as a cooperator or non-cooperator. This result is in line with research that shows that negative emotions can be displayed

in response to unfair offers in an ultimatum game (Chapman et al., 2009; Schug et al., 2010). The results were also compatible with the hypothesis that emotional expressivity can be linked to the cooperative tendency of individuals (Boone & Buck, 2003; Schug et al., 2010). The argument is that people who are emotionally expressive are likely to reveal (or leak) their motivational intentions to potential interaction partners through involuntary signals (such as facial expression of emotion). Thus, from an evolutionary perspective, non-cooperators must learn to hide their emotions or, if they cannot control their displays, commit to having cooperative intentions; otherwise, they would be avoided as interaction partners and risk extinction. Our results are compatible with this hypothesis, because they show that people are capable of identifying non-cooperators from (positive or negative) emotion displays and punish them. However, the results did not confirm (or disconfirm) the hypothesis that emotional expressivity is a reliable signal of cooperation and further research is necessary to establish, in general, that cooperators are more emotionally expressive than non-cooperators. Finally, the social-functions of emotion have also been argued to be useful in understanding the impact of emotion displays in negotiation (Morris & Keltner, 2000). Building on this perspective, Van Kleef et al. (2010) propose that the social effects of emotion can be achieved through affective or inferential processes. In the latter case, emotion displays are interpreted as information signals about the counterpart's intentions. The reverse appraisal proposal is compatible with this model in that it can be understood as a mechanism, based on appraisal variables, for such inferential processes.

5.5 Implications for Emotion Theory

The evidence presented in this dissertation emphasizes that emotions serve important social functions (Frijda & Mesquita, 1994; Keltner & Haidt, 1999; Keltner & Kring, 1998; Morris & Keltner, 2000; Oatley & Jenkins, 1996). Effectively, the results in Study 3 suggested the following social functions for emotions in the context of a social dilemma: anger punishes the

other for defecting (Morris & Keltner, 2000) and signals unwillingness to cooperate in the future; guilt appeases, serves as an apology (Keltner & Buswell, 1997; Ketelaar & Au, 2003) and signals willingness to cooperate in the future; sadness signals dissatisfaction with the current outcome, and appeals for more cooperation from the other side; a smile signals happiness with the current state of affairs, which does not necessarily mean a willingness to cooperate in the future but a willingness to maintain the same course of action. The social effects of a smile, in particular, emphasize the importance of context and are in line with previous findings suggesting the effects of emotion displays are influenced by context (Hareli & Hess, 2010; Van Kleef et al., 2010).

The dissertation proposes further that reverse appraisal is a useful framework to understand the social function of emotion of conveying information about one's intentions. The reverse appraisal proposal is that people can infer, from emotion displays, how others are appraising the situations, which in turn supports inferences about other's mental states. The first experiment in Study 3 showed that emotion displays influenced how people perceived others to be appraising the situation: a smile meant the other party found the outcome conducive to his or her goals; anger meant the other party found the outcome obstructive and blamed the participant for it; sadness meant the other party found the outcome obstructive; finally, guilt meant the other party found the outcome obstructive and blamed himself for it. Notice these patterns closely match expectations from appraisal theories (Ellsworth & Scherer, 2003). The second experiment in Study 3, in turn, showed that perceptions of appraisal had an effect on perception of how likely the other was to cooperate in the future. Moreover, this effect was similar to the effect of emotion displays in Experiment 1. Finally, Experiment 1 also showed that perception of appraisals (partially and, in some cases fully) mediated the effect of emotion displays on perception of cooperativeness. Collectively, the results, thus, provide evidence (Baron & Kenny, 1986; Spencer et al., 2005) for a causal model where, at least in the context of a social dilemma, emotion

displays cause one to perceive how the other is appraising the situation, which in turn causes perceptions about the other's likelihood of cooperation in the future. The reverse appraisal proposal, furthermore, can potentially be generalized beyond social dilemmas and be viewed as a general mechanism for the interpretation of emotion displays. This is, in fact, a promising line of future inquiry.

Finally, the article also makes a methodological contribution for the study of emotion. First, in line with the idea that virtual technology and virtual agents can be used for basic social psychology research (Blascovich et al., 2002), all studies used virtual agents to create the experimental manipulations. Using virtual agents allowed precise experimental control, low-cost, easy, replicable, and incremental research. Moreover, aside from extending current knowledge, the findings were compatible with (and, in some cases, replicated) previous findings from the behavioral sciences regarding human-human interaction, suggesting people interact naturally with virtual agents. Second, we propose social dilemmas are a useful domain for the study of emotion. Social dilemmas constitute a central problem in social interaction in which emotion plays a role and, support easy measurement of relevant behavioral variables. Moreover, this domain also benefits emotion research with plenty of previous findings and research methods from the extensive experimental economics literature (Hertwig & Ortmann, 2001).

5.6 Future Work

The research initiated in this dissertation has long-term goals and there is, of course, much future work ahead. Some of the shorter-term goals include:

- *Develop further the reverse appraisal proposal*. The studies presented in this dissertation explored only a subset of the existent emotions and appraisal variables. To address this limitation two challenges need to be considered: (a) basic research in appraisal theories is needed as researchers agree that the appraisal dimensions proposed so far are "neither final

nor complete" (p. 236, Mesquita & Ellsworth, 2001) and there is still divergence regarding some appraisal-emotion patterns (Ellsworth & Scherer, 2003); and, (b) cultural (Mesquita & Ellsworth, 2001) and personality (Krohne, 2003) factors should be considered as they are known to influence how people perceive emotions and appraisals;

- *Expand beyond the prisoner's dilemma*. Whereas the present research focused exclusively on the prisoner's dilemma, there is no reason not to expect emotion displays to impact cooperation rates in other two-person dilemmas (e.g., assurance game) or multiple-person dilemmas (e.g., public goods dilemma). Aside from social dilemmas, research in social conflict also focuses on negotiation as a mechanism for the resolution of divergence of interests (Pruitt & Carnevale, 1993). In negotiation there are usually multiple issues under consideration and parties must try to reach agreement on all of them. Recent research in the behavioral sciences has shown that, in fact, expression of emotions by the other party can influence one's concession-making (Van Kleef, De Dreu, & Manstead, 2004; Van Kleef, De Dreu, & Manstead, 2006; Van Kleef et al., 2010). For instance, these studies demonstrate that displaying anger in negotiation often triggers greater concession-making in one's opponent, whereas displaying happiness leads to fewer concessions. Recently, we have shown that these findings also carry to human-agent interaction (de Melo, Carnevale, & Gratch, 2011a). A limitation in all these studies is that the expression of emotion is scripted and not-contingent on the offers participants make. However, non-contingent display of emotion is at odds with appraisal theories of emotion. For instance, if the opponent displays anger when the participant makes a bad offer, people can infer that the opponent does not like the offer and is blaming them for that. However, what does it mean when the opponent expresses anger and the offer was good? Following the insight gained in this dissertation, it is worthwhile

exploring whether the context in which emotion is expressed impacts the interpretation of emotion and, thus, negotiation outcome;

- *Having made the inference from emotion displays about the agent's propensity for cooperation, when will people actually cooperate?* Study 3 argued for a causal model where emotion displays cause perceptions about how the agent is appraising the interaction which, in turn, cause perceptions about the agent's likelihood of cooperation. However, the link from perceptions of cooperation to the actual decision to cooperate requires further study. Research has shown that people with different social value orientations (e.g., pro-social or pro-self) act differently when faced with a cooperator (McClintock & Liebrand, 1988). Selfish individuals tend to behave non-cooperatively, whereas pro-social individuals tend to trust others and cooperate more. However, when pro-socials can identity a competitor, they are capable of adapting their behavior and punish with levels of cooperation that are even lower than that of non-cooperators (Steinel & de Dreu, 2004). Other individual factors have also been argued to impact how people behave in conflict (Christie & Geis, 1970; Rotter, 1980; Thomas, 1976). Aside from dispositional traits, situational aspects also bear influence on cooperation (Frank, 2004; Sally, 2000). For instance, Studies 1 and 2 reveal that the order with which people play a cooperator and non-cooperator impacts cooperation. There are, therefore, many factors that might influence whether a person, having identified the agent's intentions, will decide to cooperate;
- *Explore further computer models of decision-making in social dilemmas*. The current models still have relatively high error rates (see, for instance, Table 4.2 or Table 4.9). This might reflect that important features that characterize how people decide in social dilemmas are missing. The proposed models are also only a first step towards a broader domain-independent reverse appraisal model that supports general interpretation of emotion displays.

This model would nicely complement existent computer models of emotion based on appraisal theory (Marsella et al., 2010), such as EMA (Marsella & Gratch, 2004) which is a domain-independent model that integrates appraisal processes into a belief-desire-intention (BDI) architecture. Moreover, such a model would capture the abstract function of emotion displays and, thus, could be used with systems that extend far beyond virtual agents. Finally, in order to approximate this research with current investigation in game-theoretic models in artificial intelligence, it would be interesting to align the reverse appraisal model with current models in game theory that allow communication between players (for a recent survey see: Forges, 2009);

- *Understand whether people are, aside from interpreting the information conveyed by emotion displays, experiencing emotions themselves*. Keltner and Kring (1998) argue that emotion serves several social functions including the *informative* and *evocative* functions. The informative function argues that emotions convey information about the person's intentions, desires and beliefs. Our current results provide ample empirical support for this function of emotion. The evocative function, on the other hand, means that emotion displays can evoke complementary emotions in others (e.g., if someone is shown anger, the recipient responds with fear). Alternatively, Hatfield et al. (1994) argue that people can, in some cases, "catch" others' emotions (e.g., displays of anger in others lead oneself to also experience anger). The studies reported in the dissertation, however, do not clarify whether emotions are being elicited in the participants. The broader question is: are participants simply taking the information conveyed in emotion displays and deciding in a "cold" state, or are they getting emotionally aroused? Though the answer to this question can be probed with self-report measures, the best option is to also have *physiological* measures. The latter have the advantage of being measured at the exact time instant the stimulus occurs and can reflect

processes that are unconscious. Moreover, there is research exploring how different appraisal patterns result in different physiological patterns (Blascovich & Mendes, 2010; Pecchinenda, 2001). Thus, if participants are aroused distinctively by emotion displays, then we would also be able to compare the physiological patterns with expected patterns for appraisal variables. This would, potentially, constitute further evidence for the reverse appraisal proposal.

5.7 Conclusion

Modern technology has made possible the creation of computers that take on some of the social responsibilities that were usually expected only of humans. We have shown in this dissertation how computer programs that simulate expression of emotion can impact people's decision-making and promote cooperation. As computing becomes more ubiquitous in life, it is critical that computer systems of the future have this kind of emotional competency. Moreover, this dissertation emphasizes the importance of creating a solid foundation for such systems based on psychological theories of human behavior. Not only can such multi-disciplinary research inform the design of socially intelligent computer systems but, it can also facilitate the development of the underlying psychological theories. In this dissertation, virtual human technology was used to develop the reverse appraisal theory that explains how people retrieve information about appraisals from emotion displays which, in turn, informs inferences about the counterpart's mental state. On the other hand, reverse appraisal feeds back to the design of effective and natural human-interaction systems by characterizing the abstract function of emotion displays. According to Herbert Simon (1969), this is what we need to learn from nature and replicate in artificial systems. In the same way emotion evolved to endow humans with a quick and effective mechanism to solve recurrent problems that occur in social interaction, a similar mechanism is necessary to solve recurrent problems in evermore complex human-computer or computer-computer interaction systems. Emotion can be such a mechanism.

Bibliography

Akenine-Moller, T., Haines, E., & Hoffman, N. (2008). *Real-time rendering*, 3rd edition. Natick, MA: A. K. Peters, Ltd.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'ecole Americaine. *Econometrica*, *21*(4), 503-546.

Alpaydin, E. (2010). *Introduction to machine learning*, 2nd edition. Cambridge, MA: MIT Press.

Arrow, K. (1971). *Essays in the theory of risk-bearing*. Chicago, IL: Markham.

Aumann, R. (1997). Rationality and bounded rationality. *Games and Economic Behavior*, *21*(1-2), 2-14.

Averill, J. (1980). A constructivist view of emotion. In R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory research, and experience* (pp. 305-339). Orlando, FL: Academic Press.

Axelrod, R. (1984). *The evolution of cooperation*. New York, NY: Basic Books.

Balliet, D. (2010). Communication and cooperation in social dilemmas: A meta-analytic review. *Journal of Conflict Resolution*, *54*(1), 39-57.

Baron, R., & Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173-1182.

Baumeister, R., Heatherton, T., & Tice, D. (1994). *Losing control: How and why people fail at self-regulation*. San Diego, CA: Academic Press.

Bavelas, J., Black, A., Lemery, C., & Mullet, J. (1986). 'I show how you feel': Motor mimicry as a communicative act. *Journal of Personality and Social Psychology, 50*(2), 322-329.

Bazerman, M., Tenbrunsel, T., Wade-Benzoni, K. (1998). Negotiating with yourself and losing: Understanding and managing conflicting internal preferences. *Academy of Management Review, 23*(2), 225-241.

Beale, R., & Creed, C. (2009). Affective interaction: How emotional agents affect users. *Human-Computer Studies, 67*(9), 755-776.

Bechara, A., Damasio, A., Damasio, H., & Anderson, S. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*(1-3), 7-15.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. (1997). Deciding advantageously before knowing the advantageous strategy. *Science*, *275*(5304), 1293-1294.

Becker-Asano, C., & Wachsmuth, I. (2008). Affect simulation with primary and secondary emotions. In *Proceedings of the 8th International Conference on Intelligent Virtual Agents.*

Bente, G., Kramer, N., Petersen, A., & de Ruiter, J. (2001). Computer animated movement and person perception: Methodological advances in nonverbal behavior research. *Journal of Nonverbal Behavior*, *25*(3), 151-166.

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. (original 1738) *Econometrica*, *22*(1), 23-36.

Berry, D., Butler, L., & De Rosis, F. (2005). Evaluating a realistic agent in an advice-giving task. *Journal of Human-Computer Studies*, *63*(3), 304-327.

Blanchette, I., & Campbell, M. (2005). The effect of emotion on syllogistic reasoning in a group of war veterans. Paper presented at the *27th Annual Conference of the Cognitive Science Society*, Stresa, Italy.

Blanchette, I., & Richards, A. (2004). Reasoning about motional and neutral materials. Is logic affected by emotion? *Psychological Science*, *15*(11), 745-752.

Blanchette, I., & Richards, A. (2010). The influence of affect on higher level cognition: A review of research on interpretation, judgment, decision making and reasoning. *Cognition & Emotion*, *15*(4), 1-35.

Blanchette, I., Richards, A., & Cross, A. (2007). Anxiety and the interpretation of ambiguous facial expressions: The influence of contextual cues. *Quarterly Journal of Experimental Psychology*, *60*(8), 1101-1115.

Blanchette, I., Richards, A., Melnyk, L., & Lavda, A. (2007). Reasoning about emotional contents following shocking terrorist attacks: A tale of three cities. *Journal of Experimental Psychology: Applied*, *13*(1), 47-56.

Blascovich, J. (2002). Social influence within immersive virtual environments. In R. Schroeder (Ed.), *The social life of avatars: Presence and interaction in shared virtual environments* (pp. 127-145). London, UK: Springer-Verlag.

Blascovich, J., & Mendes, W. (2010). Social psychophysiology and embodiment. In S. Fiske, D. Gilbert & G. Lindzey (Eds.), *Handbook of Social Psychology*, 5th edition (pp.194-226). New Jersey, NJ: John Wiley & Sons, Inc.

Blascovich, J., Mendes, W., Hunter, S. B., & Salomon, K. (1999). Social "facilitation" as challenge and threat. *Journal of Personality and Social Psychology*, *77*(1), 68-77.

Blascovich, J., Loomis, J., Beall, A., Swinth, K., Hoyt, C., & Bailenson, J. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry*, *13*(2), 103-124.

Bixenstine, V., & Wilson, K. (1963). Effects of level of cooperative choice by the other player on choices in a prisoner's dilemma game, Part II. *Journal of Abnormal and Social Psychology, 67*(2), 139-147.

Boiten, F., Frijda, N., & Wientjes, C. (1994). Emotions and respiratory patterns: Review and critical analysis. *International Journal of Psychophysiology, 17*(2), 103-128.

Boone, R., & Buck, R. (2003). Emotional expressivity and trustworthiness: The role of nonverbal behavior in the evolution of cooperation. *Journal of Nonverbal Behavior, 27*, 163-182.

Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *Journal of Human-Computer Studies, 62*(2), 161-178.

Brown, W., Palameta, B., & Moore, C. (2003). Are there nonverbal cues to commitment? An exploratory study using the zero-acquaintance video presentation paradigm. *Evolutionary Psychology, 1*, 42-69.

Camerer, C. (1995). Individual decision making. In J. Kagel & A. Roth (Eds.), *Handbook of Experimental Economics*, Princeton, NJ: Princeton University Press.

Calvo, R., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing 1*(1), 16-37.

Chapman, H., Kim, D., Susskind, J., & Anderson, A. (2009). In bad taste: Evidence for the oral origins of moral disgust. *Science, 323*(5918), 1222–1226.

Christie, R., & Geis, F. (1970). *Studies in Machiavellianism*. New York, NY: Academic Press.

Creed, C., & Beale, R. (2008). Psychological responses to simulated displays of mismatched emotional expressions. *Interacting with Computers, 20*(2), 225–239.

Damasio, A. (1994). *Descartes' error: Emotion, reason and the human brain*. New York, NY: Putnam.

Dehn, D. & Van Mulken, S. (2000). The impact of animated interface agents: a review of empirical research. *International Journal of Human-Computer Studies 52*(1), 1-22.

Deutsch, M., Epstein, Y., Canavan, D., & Gumpert, P. (1967). Strategies for inducing cooperation: An experimental study. *Journal of Conflict Resolution, 11*(3), 104-109.

de Graaf, V. (2002). *Human anatomy*, 6th edition. McGraw Hill.

de Jong, P., Peters, M., de Cremer, D., & Vranken, C. (2002). Blushing after a moral transgression in a prisoner's dilemma game: appeasing or revealing? *European Journal of Social Psychology, 32*(5), 627-644.

de Melo, C. (2006). *Gesticulation expression in virtual humans*. M.Sc. Thesis. Department of Information Systems and Computer Engineering, IST–Technical University of Lisbon.

de Melo, C., & Gratch, J. (2009a). Expression of emotions using wrinkles, blushing, sweating and tears. In *Proceedings of the Intelligent Virtual Agents 2009*, pp. 188-200.

de Melo, C., & Gratch, J. (2009b). Expression of emotions using wrinkles, blushing, sweating and tears. In *Proceedings of the Intelligent Virtual Agents (IVA) 2009*, pp. 188-200.

de Melo, C., & Paiva, A. (2006a). Multimodal expression in virtual humans. *Computer Animation and Virtual Worlds*, *17*, 239-248.

de Melo, C., & Paiva, A. (2006b). A story about gesticulation expression. In *Proceedings of the Intelligent Virtual Agents Conference (IVA) 2006*.

de Melo, C., & Paiva, A. (2007). Expression of emotions in virtual humans using lights, shadows, composition and filters. In *Proceedings of the Affective Computing and Intelligent Interaction (ACII) Conference 2007*.

de Melo, C., & Paiva, A. (2008a). Evolutionary expression of emotions in virtual humans using lights and pixels. In J. Tao, & T.N. Tan (Eds.), *Affective Information Processing* (pp.313–336). Springer Science+Business Media LLC.

de Melo, C., & Paiva, A. (2008b). Modeling gesticulation expression in virtual humans. In N. Magnenat-Thalmann, L. Jain, & N. Ichalkaranje (Eds.), *New advances in virtual humans: Artificial intelligence environment* (pp.133-151). Berlin, Germany: Springer-Verlag.

de Melo, C., Carnevale, P., & Gratch, J. (2010). The influence of emotions in embodied agents on human decision-making. In *Proceedings of Intelligent Virtual Agents (IVA) 2010*, pp. 357-370.

de Melo, C., Carnevale, P., & Gratch, J. (2011a). The effect of expression of anger and happiness in computer agents on negotiations with humans. In *Proceedings of Autonomous Agents and Multiagent Systems (AAMAS) 2011*.

de Melo, C., Carnevale, P., & Gratch, J. (2011b). Reverse appraisal: Inferring from emotion displays who is the cooperator and the competitor in a social dilemma. In *Proceedings of The 33rd Annual Meeting of the Cognitive Science Society (CogSci'11)*.

de Melo, C., Carnevale, P., Gratch, J. (2012). The impact of emotion displays in embodied agents on emergence of cooperation with people. *Presence: Teleoperators and Virtual Environments Journal*, *20*(5), 449-465.

de Melo, C., Kenny, P., & Gratch, J. (2010a). The influence of autonomic signals on perception of emotions in embodied agents. *Applied Artificial Intelligence*, *24*(6), 494-509.

de Melo, C., Kenny, P., & Gratch, J. (2010b). Real-time expression of affect through respiration. *Computer Animation and Virtual Worlds*, *21*, 225-234.

de Melo, C., Zheng, L., & Gratch, J. (2009). Expression of moral emotions in cooperating agents. In *Proceedings of Intelligent Virtual Agents (IVA) 2009*, pp. 301-307.

de Melo, C., Carnevale, P., Antos, D., & Gratch, J. (2011). A computer model of the interpersonal effect of emotion displayed in a social dilemma. In *Proceedings of Affective Computing and Intelligent Interaction (ACII) 2011*.

de Melo, C., Carnevale, P., Read, S., Antos, D., & Gratch, J. (2012). Bayesian Model of the Social Effects of Emotion in Decision-Making in Multiagent Systems. Paper to be presented at the *11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Valencia, Spain.

Dias, J., & Paiva, A. (2005). Feeling and reasoning: A computational model for emotional agents. In *Proceedings of the 12th Portuguese Conference on Artificial Intelligence, EPIA 2005*.

Doyle, J. (1998) Rational decision making. In R. Wilson & F. Kiel (Eds.), *The MIT Encyclopedia of the Cognitive Sciences* (pp.30-34). Cambridge, MA: MIT Press.

Ekman, P. (1999a). Basic emotions. In T. Dalgleish, & M. Power (Eds.), *Handbook of Cognition and Emotion* (pp.45-60). New York, NY: John Wiley & Sons Ltd.

Ekman, P. (1999b). Facial expressions. In T. Dalgleish, & M. Power (Eds.), *Handbook of Cognition and Emotion* (pp.281-300). New York, NY: John Wiley & Sons Ltd.

Ellsworth, P., & Scherer, K. (2003). Appraisal processes in emotion. In R. Davidson, K. Scherer, H. Goldsmith (Eds.), *Handbook of Affective Sciences* (pp. 572-595). New York, NY: Oxford University Press.

Fernandez-Dols, J., & Ruiz-Belda, M. (1995). Are smiles signs of happiness? Gold medal winners at the Olympic games. *Journal of Personality and Social Psychology*, *69*(6), 1113-1119.

Forges, F. (2009). Correlated equilibria and communication in games. In R. Meyers (Ed.), *Encyclopedia of Complexity and Systems Science*, Springer.

Frank, R. (1988). *Passions within reason.* New York, NY: Norton.

Frank, R. (2004). Introducing moral emotions into models of rational choice. In A. S. R. Manstead, N. Frijda & A. Fischer (Eds.), *Feelings and emotions* (pp. 422-440). New York, NY: Cambridge University Press.

Friedman, M., & Savage, L. (1948). The utility analysis of choices involving risks. *Journal of Political Economy*, *56*(4), 279-304.

Frijda, N. (1986). *The emotions*. Cambridge, UK: Cambridge University Press.

Frijda, N. (1989). Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology*, *57*, 212-228.

Frijda, N., & Mesquita, B. (1994). The social roles and functions of emotions. In S. Kitayama & H. Markus (Eds.), *Emotion and culture: Empirical studies of mutual influence* (pp. 51–87). Washington, DC: American Psychological Association.

Gong, L. (2007). Is happy better than sad even if they are both non-adaptive? Effects of emotional expressions of talking–head interface e-agents. *Journal of Human-Computer Studies*, *65*(3), 183-191.

Gratch, J., & Marsella, S. (2004). A domain independent framework for modeling emotion. *Journal of Cognitive Systems Research*, *5*(4), 269-306.

Gratch, J., Rickel, J., Andre, E., Badler, N., Cassell, J., & Petajan, E. (2002). Creating interactive virtual humans: Some assembly required. *IEEE Intelligent Systems*, *17*(4), 54-63.

Griffiths, T., Kemp, C., & Tenenbaum, J. (2008). Bayesian models of cognition. In R. Sun, (Ed.), *The Cambridge handbook of computational cognitive modeling*. Cambridge, MA: Cambridge University Press.

Hareli, S., & Hess, U. (2010). What emotional reactions can tell us about the nature of others: An appraisal perspective on person perception. *Cognition & Emotion*, *24*(1), 128-140.

Harford, T., & Solomon, L. (1967). 'Reformed sinner' and 'lapsed saint' strategies in the prisoner's dilemma game. *Journal of Conflict Resolution*, *11*(1), 345-360.

Hatfield, E., Cacioppo, J., & Rapson, R. (1994). *Emotional contagion*. New York, NY: Cambridge University Press.

Hazan, C., & Shaver, P. (1987). Romantic love conceptualized as an attachment process. *Journal of Personality and Social Psychology*, *52*(3), 511-524.

Heckerman, D., Geiger, D., & Chickering, D. (1995). Learning Bayesian Networks: The combination of knowledge and statistical data. *Machine Learning*, *20*(3), 197-243.

Helson, H. (1964). *Adaptation-level theory*. New York, NY: Harper & Row.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, *24*(3), 383-451.

Hirschman, A. (1997). *The passions and the interests*. Princeton, NJ: Cambridge University Press.

Hilty, J., & Carnevale, P. (1993) Black-hat/white-hat strategy in bilateral negotiation. *Organizational Behavior and Human Decision Processes*, *55*(3), 444-469.

Hone, K. (2006). Empathic agents to reduce user frustration: The effects of varying agent characteristics. *Interacting with Computers*, *18*(2), 227-245.

Jerdee, T., & Rosen, B. (1974). Effects of opportunity to communicate and visibility of individual decisions on behavior in the common interest. *Journal of Applied Social Psychology*, *59*(6), 712-16.

Johnson-Laird, P., & Oatley, K. (1992). Basic emotions, rationality, and folk theory. *Cognition and Emotion*, *6*(3), 201-223.

Keeney, R., & Raiffa, H. (1976). *Decisions with multiple objectives: Preferences and value tradeoffs*. New York, NY: Wiley.

Keltner, D., & Buswell, B. (1997). Embarrassment: Its distinct form and appeasement functions. *Psychological Bulletin*, *122*(3), 250-270.

Keltner, D., & Ekman, P. (2000). Facial expression of emotion. In M. Lewis & J. Haviland-Jones (Eds.), *Handbook of Emotion* (pp. 236-249). New York, NY: Guilford Press.

Keltner, D., & Haidt, J. (1999). Social functions of emotions at four levels of analysis. *Cognition and Emotion*, *13*(5), 505–521.

Keltner, D., & Kring, A. M. (1998). Emotion, social function, and psychopathology. *Review of General Psychology*, *2*(3), 320–342.

Kerr, N. (1989). Illusions of efficacy: The effects of group size on perceived efficacy in social dilemmas. *Journal of Experimental Social Psychology*, *25*(4), 287-313.

Kerr, N. (2011). Social dilemmas. In J. Levine (Ed.), *Frontiers of Social Psychology: Group Processes*. New York, NY: Psychology Press.

Ketelaar, T., & Au, W. (2003). The effects of feelings of guilt on the behavior of uncooperative individuals in repeated social bargaining games: An affect-as-information interpretation of the role of emotion in social interaction. *Cognition and Emotion*, *17*(3), 429-453.

Kiesler, S., Waters, K., & Sproull, L. (1996). A prisoner's dilemma experiment on cooperation with human-like computers. *Journal of Personality and Social Psychology, 70*(1), 47-65.

Klein, J., Moon, Y., & Picard, R. (2002). This computer responds to user frustration: theory, design, and results. *Interacting with Computers 14*(2), 119–140.

Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology, 24*(1), 183-214.

Kollock, P. (1998b). Transforming social dilemmas: Group identity and cooperation. In Danielson, P. (Ed.), *Modeling Rational and Moral Agents* (pp. 186-210). Oxford, UK: Oxford Univ. Press.

Komorita, S., & Parks, C. (1994). *Social dilemmas*. Dubuque, IA: Brown and Benchmark.

Kramer, R., & Brewer, M. (1986). Social group identity and the emergence of cooperation in resource conservation dilemmas. In H. Wilke, D. Messick & C. Rutte (Eds.), *Experimental Social Dilemmas* (pp. 205-34). Frankfurt: Verlag Peter Lang.

Kraus, S. (1997). Negotiation and cooperation in multi-agent environments. *Artificial Intelligence*, *94*(1-2), 79-98.

Kraut, R., & Johnston, R. (1979). Social and emotional messages of smiling: An ethological approach. *Journal of Personality and Social Psychology*, *37*(9), 1539-1533.

Krohne, W. (2003). Individual differences in emotional reactions and coping. In R. Davidson, K. Scherer, H. Goldsmith (Eds.), *Handbook of Affective Sciences* (pp. 698-725). New York, NY: Oxford University Press.

Krumhuber, E., Manstead, A., & Kappas, A. (2007). Facial dynamics as indicators of trustworthiness and cooperative behavior. *Emotion*, *7*(4), 730-735.

Kuno, Y. (1956). *Human perspiration.* Springfield: Charles C. Thomas.

Lazarus, R. (1991). *Emotion and adaptation.* New York, NY: Oxford University Press.

Leary, M., Britt, T., Cutlip, W., & Templeton, J. (1992). Social blushing. *Psychological Bulletin, 112*(3), 446-460.

Lefford, A. (1946). The influence of emotional subject matter on logical reasoning. *Journal of General Psychology*, *34*, 127-151.

Lester, J., Converse, S., Kahler, S., Barlow, T., Stone, B., & Bhogal, R. (1997). The persona effect: affective impact of animated pedagogical agents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM Press, 359-366.

Levenson, R. (2003). Autonomic specificity and emotion. In R. J. Davidson, K. R. Scherer & H. H. Goldsmith (Eds.), *Handbook of Affective Sciences* (pp.212-224). New York, NY: Oxford University Press.

Lewis, M. (2008). Self-conscious emotions: Embarrassment, pride, shame, and guilt. In L. Michael, & J. Haviland-Jones (Eds.) *Handbook of Emotions*, 623-636. New York, NY: The Guilford Press.

Lim, Y. & Aylett, R. (2007). Feel the difference: a guide with attitude!. In *Proceedings of the 7th International Conference on Intelligent Virtual Agents*, 317-330.

Lin, R., & Kraus, S. (2010). Can automated agents proficiently negotiate with humans? *Communications of the ACM*, *53*(1), 78-88.

Liu, K., & Picard, R. (2005). Embedded empathy in continuous, interactive health assessment. In *Computer–Human Interaction Workshop on Computer-Human Interaction Challenges in Health Assessment.*

Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes, 65*(3), 272-292.

Loewenstein, G., & Lerner, J. (2003). The role of affect in decision making. In R. Davidson, K. Scherer & H. Goldsmith (Eds.), *Handbook of Affective Sciences* (pp. 619-642). New York, NY: Oxford University Press.

Lutz, C., & White, G. (1986). The anthropology of emotions. *Annual Review of Anthropology, 15*(1), 405-436.

Manstead, A., & Fischer, A. (2001). Social appraisal: The social world as object of and influence on appraisal processes. In K. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 221-232). Oxford, UK: Oxford University Press.

Marsella, S., Gratch, J., & Petta, P. (2010) Computational models of emotion. In K. Scherer., T. Bänziger, & E. Roesch (Eds.), *A blueprint for an affectively competent agent: Cross-fertilization between Emotion Psychology, Affective Neuroscience, and Affective Computing*. New York, NY: Oxford University Press.

Matsumoto, D., Haan, N., Gary, Y., Theodorou, P., & Cooke-Carney, C. (1986). Preschoolers' moral actions and emotions in prisoner's dilemma. *Developmental Psychology*, *22*(5), 663-670.

McClintock, C., & Liebrand, W. (1988). Role of interdependence structure, individual value orientation, and another's strategy in social decision making: a transformational analysis. *Journal of Personality and Social Psychology*, *55*(3), 396-409.

Mehu, M., Grammer, K., & Dunbar, R. (2007). Smiles when sharing. *Evolution and Human Behavior*, *28*(6), 415-422.

Melton, R. (1995). The role of positive affect in syllogism performance. *Personality & Social Psychology Bulletin*, *21*(8), 788-794.

Mesquita, B, & Ellsworth, P. (2001). The role of culture in appraisal. In K. Scherer, A. Schorr & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 233-248). New York, NY: Oxford University Press.

Messick, D., Wilke, H., Brewer, M., Kramer, R., Zemke, P., & Lui, L. (1983). Individual adaptations and structural change as solutions to social dilemmas. *Journal of Personality and Social Psychology*, *44*(2), 294-309.

Miceli, M., & Castelfranchi, C. (2003). Crying: Discussing its basic reasons and uses. *New Ideas in Psychology*, *21*(3), 247-273.

McGregor, I. (1952). The sweating reactions of the forehead. *Journal of Physiology*, *116*(1), 26-34.

Miller, R., & Leary, M. (1992). Social sources and interactive functions of emotion: The case of embarrassment. In M. Clark (Ed.), *Emotion and Social Behavior* (pp. 202-221). Beverly Hills, CA: Sage.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago, IL: University of Chicago Press.

Minsky, M. (1986). *The society of mind*. New York, NY: Simon and Schuster.

Morris, M., & Keltner, D. (2000). How emotions work: An analysis of the social functions of emotional expression in negotiations. *Research in Organizational Behavior*, *22*, 1-50.

Nass, C., Fogg, B., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies*, *45*(6), 669-678.

Nass, C., Steuer, J., & Tauber, E. (1994). Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Nesse, R. (1990). Evolutionary explanations of emotions. *Human Nature*, *1*(3), 261-289.

Nisbett, R., & Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 14*(3), 231-259.

Oaksford, M., Morris, F., Grainger, B., & Williams, J. (1996). Mood, reasoning, and central executive processes. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *22*(2), 476-492.

Oatley, K., & Jenkins, J. (1996). *Understanding emotions*. Cambridge, MA: Blackwell.

Orbell, J., van de Kragt, A., & Dawes, R. (1988). Explaining discussion-induced cooperation. *Journal of Personality and Social Psychology*, *54*(5), 811-819.

Ortony, A., Clore, G., & Collins, A. (1988). *The cognitive structure of emotions.* New York, NY: Cambridge University Press.

Osborne, M., & Rubinstein, A. (1994). *A course in game theory*. Cambridge, MA: MIT Press.

Palfai, T., & Salovey, P. (1993). The influence of depressed and elated mood on deductive and inductive reasoning. Imagination. *Cognition and Personality*, *13*, 57-71.

Pantic, M., Pentland, A., Nijholt, A. & Huang, T. (2006). Human computing and machine understanding of human behavior: A survey. *Proceedings of the 8th ACM International Conference on Multimodal Interfaces (ICMI'06)*. 239-248.

Parke, F. (1972). Computer generated animation of faces. In *Proceedings of SIGGRAPH 1972*.

Parke, F. (1974). *A parametric model for human faces*. PhD thesis. University of Utah.

Pecchinenda, A. (2001). The Psychophysiology of Appraisals. In K. Scherer, A. Schorr & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 301-315). New York, NY: Oxford University Press.

Picard, R. (1997). *Affective computing*. Cambridge, MA: The MIT Press.

Poundstone, W. (1993). *Prisoner's dilemma*. New York, NY: Doubleday.

Preacher, K., & Hayes, A. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*(3), 879-891.

Prendinger, H., Mayer, S., Mori, J., Ishizuka, M. (2003). Persona effect revisited. Using bio-signals to measure and reflect the impact of character-based interfaces. In *Proceedings of Fourth International Working Conference On Intelligent Virtual Agents* (IVA03), 283-291.

Pruitt, D., & Carnevale. (1993). *Negotiation in social conflict*. Pacific Grove, CA: Brooks/Cole.

Pruitt, D., & Kimmel, M. (1977). Twenty years of experimental gaming: Critique, synthesis, and suggestions for the future. *Annual Review of Psychology, 28*, 363-392.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. New York, NY: Cambridge University Press.

Roseman, I. (2001). A model of appraisal in the emotion system: Integrating theory, research, and applications. In K. Scherer, A. Schorr & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 68-91). New York, NY: Oxford University Press.

Roseman, I., & Smith, C. (2001). Appraisal theory: Overview, assumptions, varieties, controversies. In K. Scherer, A. Schorr & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 3-19). New York, NY: Oxford University Press.

Roseman, I., & Spindel, M. (1990). Appraisals of emotion-eliciting events: Testing a theory of discrete emotions. *Journal of Personality and Social Psychology, 59*, 899-915.

Ross, W., & LaCroix, J. (1996). Multiple meanings of trust in negotiation theory and research: A literature review and integrative model. *International Journal of Conflict Management, 7*(4), 314-360.

Rotter, J. (1980). Interpersonal trust, trustworthiness, and gullibility. *American Psychologist, 35*(1), 1-7.

Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach*, 3rd edition. New Jersey, NJ: Pearson.

Sally, D. (2000). A general theory of sympathy, mind-reading, and social interaction, with an application to the prisoner's dilemma. *Social Science Information, 39*(4), 567-634.

Sayre, H. (2007). *A world of art*, 5th edition. New Jersey, NJ: Prentice Hall.

Scharlemann, J., Eckel, C., Kacelnik, A., & Wilson, R. (2001). The value of a smile: Game theory with a human face. *Journal of Economic Psychology*, *22*(5), 617-640.

Scherer, K. (2001). Appraisal considered as a process of multi-level sequential checking. In K. Scherer, A. Schorr & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 92-120). New York, NY: Oxford University Press.

Scherer, K., & Grandjean, D. (2008). Facial expressions allow inference of both emotions and their components. *Cognition and Emotion*, *22*(5), 789-801.

Scheutz, M., & Schermerhorn, P. (2009). Affective goal and task selection for social robots. In J. Vallverdú & D. Casacuberta (Eds.), *The Handbook of Research on Synthetic Emotions and Sociable Robotics* (pp. 74-87). IGI Global.

Scheutz, M., & Sloman, A. (2001). Affect and agent control: Experiments with simple affective states. In *Proceedings of the Intelligent Agent Technology Conference* (pp. 200-209).

Schug, J., Matsumoto, D., Horita, Y., Yamagishi, T., & Bonnet, K. (2010). Emotional expressivity as a signal of cooperation. *Evolution and Human Behavior, 31*(2), 87-94.

Shoham, Y., & Leyton-Brown, K. (2008). *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. New York, NY: Cambridge University Press.

Simon, H. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*(2), 129-138.

Simon, H. (1967). Motivational and emotional controls of cognition. *Psychological Review, 74*(1), 29-39.

Simon, H. (1969). *The sciences of the artificial*. Cambridge, MA: MIT Press.

Simon, H. (1997). *Models of bounded rationality*. Cambridge, MA: MIT Press.

Sloman, A., & Croucher, M. (1981). Why robots will have emotions. Paper presented at the *International Joint Conference on Artificial Intelligence*, Vancouver, Canada.

Smith, C., & Ellsworth, P. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, *48*, 813-838.

Spencer, S., Zanna, M., & Fong, G. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology, 89*(6), 845-851.

Staller, A., & Petta, P. (2001). Introducing emotions into the computational study of social norms: A first evaluation. *Journal of Artificial Societies and Social Simulation, 4*(1).

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for descriptive theory of choice under risk. *Journal of Economic Literature, 38*(2), 332-382.

Steinel, W., & de Dreu, C. (2004). Social motives and strategic misrepresentation in social decision making. *Journal of Personality and Social Psychology, 86*(3), 419-434.

Taylor, P., Black, A., & Caley, R. (1998). In *Proceedings of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis*.

Thomas, K. (1976). Conflict and conflict management. In M. D. Dunnette (ed.), *Handbook of industrial & organizational psychology* (pp.889-935). Chicago: Rand McNally.

Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology, 46*(1), 35-57.

Tversky, A., & Kahneman, D. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263-292.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453-458.

Van Kleef, G., De Dreu, C., & Manstead, A. (2004). The interpersonal effects of anger and happiness in negotiations. *Journal of Personality and Social Psychology, 86*(1), 57-76.

Van Kleef, G., De Dreu, C., & Manstead, A. (2006). Supplication and appeasement in negotiation: The interpersonal effects of disappointment, worry, guilt, and regret. *Journal of Personality and Social Psychology, 91*(1), 124-142.

Van Kleef, G., De Dreu, C., & Manstead, A. (2010). An interpersonal approach to emotion in social decision making: The emotions as social information model. *Advances in Experimental Social Psychology, 42*(10), 45-96.

van Lange, P., Liebrand, W., Messick, D., & Wilke, H. (1992). Introduction and literature review. In W. Liebrand, D. Messick and H. Wilke (Eds.), *Social dilemmas: Theoretical issues and research findings* (pp. 3-28). Oxford, UK: Pergamon Press.

Van Mulken, S., André, E., & Muller, J. (1998). The persona effect: how substantial is it? In *People and Computers XIII: Proceedings of HCI'98*, 53–66.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. New York, NY: John Wiley and Sons.

Waters, K. (1987). A muscle model for animating three-dimensional facial expression. In *Proceedings of SIGGRAPH 1987*.

Wehrle, T., & Scherer, K. (2001). Toward computational modeling of appraisal theories. In K. Scherer, A. Schorr & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 350-365). New York, NY: Oxford University Press.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*(7), 592-598.

Wilson, T., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*, *116*(1), 117-142.

Wilson, T., & Schooler, J. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, *60*(2), 181-192.

Winett, R., Kagel, J., Battalio, R., & Winkler, R. (1978). The effects of monetary rebates, feedback, and information on residential energy conservation. *Journal of Applied Psychology*, *63*(1), 73-80.

Wooldridge, M. (2009). *An introduction to multiagent systems*, 2nd edition. West Sussex, UK: John Wiley & Sons Ltd.

Wubben, M., De Cremer, D., & van Dijk, E. (2008). When emotions of others affect decisions in public good dilemmas: An instrumental view. *European Journal of Social Psychology*, *38*(5), 823-835.

Wubben, M., De Cremer, D., & van Dijk, E. (2009). How emotion communication guides reciprocity: Establishing cooperation through disappointment and anger. *Journal of Experimental Social Psychology*, *45*(1), 987-990.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, *51*(1), 110-116.

Yee, N., Bailenson, J., & Rickertsen, K. (2007). A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces. In *Proceedings of the SIGCHI'07*.

Zeng, Z., Pantic, M., Roisman, G., & Huang, T. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence 31*(1), 39-58.

Appendix: Virtual Humans Platform

Overview

Virtual humans (or *embodied agents*) are agents that inhabit virtual worlds with three-dimensional bodies and that can express themselves through their bodies like people do (Gratch et al., 2002). Nass and colleagues (Nass, Steuer, & Tauber, 1994; Reeves & Nass, 1996) have shown empirical evidence that people can treat computers like other people. For example, when participants were placed on the same team as a computer for a task, they rated the computer more favorably than if the computer were not labeled a teammate (Nass, Fogg, & Moon, 1996). These findings suggest that people can also treat virtual humans like other people. Effectively, Blascovich (2000) shows empirical evidence and advances a theory of how people can be socially influenced by virtual humans. Moreover, Blascovich et al. (2002) argue that virtual human technology can be used as a methodological tool for basic research in social psychology. For instance, there are several advantages in using virtual humans over confederates in experimental settings: (1) Researchers have more experimental control with virtual humans. Confederates can inadvertently introduce noise, as their performance can have slight but relevant differences between participants; (2) Virtual humans can be carefully animated and tested before running the experiment, whereas confederates improvise in real-time; (3) Virtual humans are less expensive than confederates. Several empirical studies have replicated findings in the behavioral sciences while substituting people with virtual humans (e.g., Bente, Kramer, Petersen, & de Ruiter, 2001; Blascovich, Mendes, Hunter, & Salomon, 1999). Using the virtual humans platform described here, we also replicated a finding about how people negotiate with human adversaries that display

emotion (Van Kleef, De Dreu, & Manstead, 2004), but substituting the adversaries for virtual humans (de Melo, Carnevale, & Gratch, 2011a).

This Appendix focuses on the relevant aspects of the virtual humans platform used in the empirical studies reported in the dissertation. However, broadly, the platform integrates and synchronizes several expression modalities (for a detailed overview see: de Melo & Paiva, 2006a; de Melo, 2006). Regarding bodily expression, the platform supports keyframe animation of the skeleton, inverse kinematics to animate the limbs based on the position of the hands, and animation of gestures based on features such as hand shape, position and orientation (de Melo, 2006; de Melo & Paiva, 2006b; de Melo & Paiva, 2008). The platform supports real-time integration with the Festival speech synthesizer (Taylor, Black, & Caley, 1998), lip synching and, in order to support the kind of speech-gesture integration we see in daily conversation (McNeill, 1992), sub-second synchronization with gestures. Respiration has also been shown to reflect emotions (Boiten, Frijda, & Wientjes, 1994) and the platform supports animation of specific respiration patterns for several emotions (de Melo, Kenny, & Gratch, 2010a; de Melo, Kenny, & Gratch, 2010b). Inspiring on the idea that artists can tell stories using acting, motion, color, scenarios and lighting (Sayre, 2007), the platform also explores integration with channels of expression in the virtual human's surrounding environment such as lighting and cameras (de Melo & Paiva, 2007; de Melo & Paiva, 2008a; de Melo & Gratch, 2009a). Finally, facial expression is based on deformation of muscles in the face. Wrinkles and blushing of the face are supported as well (de Melo & Gratch, 2009b). Since facial expression is extensively used in the empirical studies described in the dissertation, the rest of the Appendix focuses on it.

Pseudo-Muscular Model for Facial Expression

The pioneering work of Parke (1972) introduced basic polygon-based face modeling and keyframe animation. The immense variety of facial expressions made this approach extremely

data-intensive, and prompted the development of parametric models (Parke, 1974). Here, expressions were created by assigning values to parameters that control facial geometry subsets. The limitations of ad hoc parametric solutions led to anatomical-based parametric models. Waters (1987) proposed a widely cited parameterized muscle-based model that simulated several kinds of muscles including *linear/parallel* muscles, that contract longitudinally from a static bone attachment towards the other end embedded in the soft tissue of the skin. Our platform implements Waters equations for deforming muscles. *Atomic* parameters control the deformation of each muscle in a standardized range of [0.0, 1.0], being 0.0 no deformation and, 1.0 maximum deformation. Thirty-seven atomic parameters are defined, each corresponding to a muscle in the human face (de Graaf, 2002). *Skinning* parameters are defined to control rotation of the eyes, jaw and tongue. Finally, *group* parameters aggregate the effect of many atomic or skinning parameters. A tool was developed to edit all these parameters (Figure A.1).

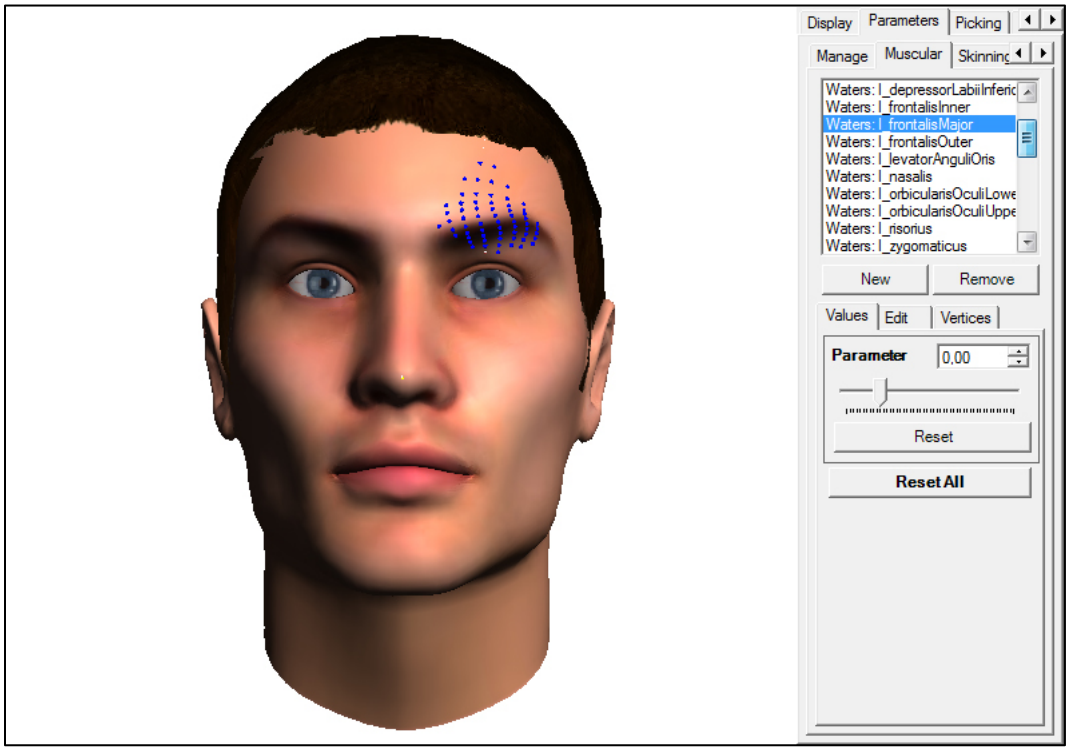


Figure A.1. Software to edit the virtual human face muscle model.

Expression of Emotions

Group parameters are used to define facial expressions of emotion. The emotion expressions used in this dissertation are defined according to Ekman's basic emotions specification (Ekman, 1999a; Ekman, 1999b). For instance, anger is accomplished by the composite effect of the following muscles, Figure A.2: raising of the (right and left) corrugator supercilli; partial closing of the lower orbicularis oculi; and, lowering of the depressor anguli oris. However, group parameters only define the muscular configuration of the emotional expression. Other physiological aspects of the facial expression–such as wrinkles or blushing–are described below in the section "Blushing, Wrinkles, Sweating and Tears".

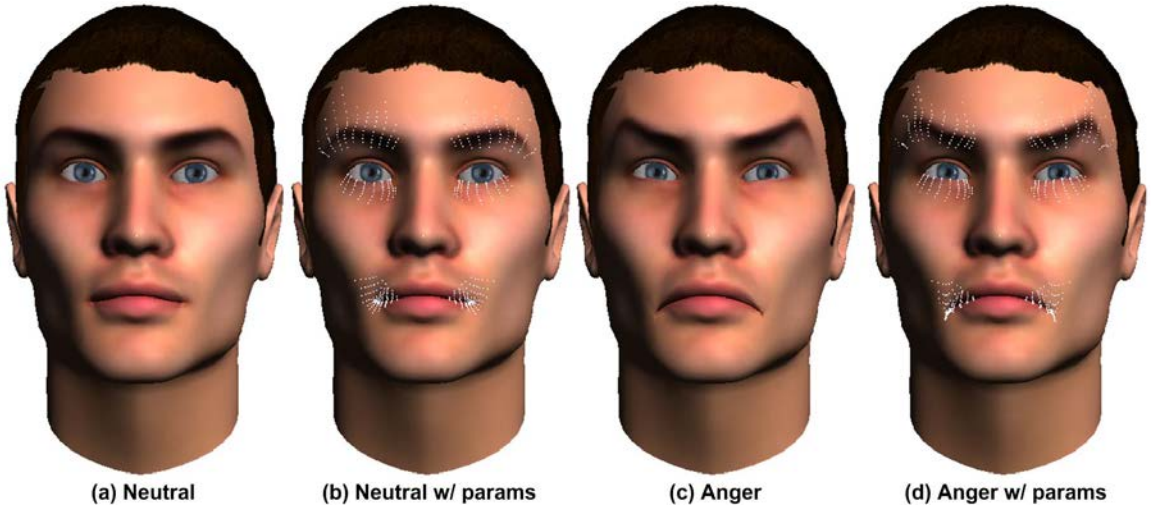


Figure A.2. Facial muscle configuration for anger.

Integration with FaceGen

Researchers have previously argued that using photorealistic virtual humans (vs. non-realistic virtual humans) can impact human-computer interaction (Blascovich, 2002; Yee, Bailenson, & Rickertsen, 2007). Thus, starting with our second empirical study, we integrated FaceGen[6]–a tool

[6] http://www.facegen.com/

that supports creation of photorealistic three-dimensional faces–with the virtual human platform. This integration consisted of a tool to import a FaceGen face into the platform's virtual human format and, another tool to edit the virtual human's "props" (hair, teeth and eyes). Once the face is imported, it becomes possible to animate it with the pseudo-muscular model described in the previous section. Moreover, since expressions (such as emotions) are defined at the muscle-level, they can easily be reused with new faces. With this pipeline, new faces can be created and animated using the platform very quickly. Overall, several new faces (Figure A.3) have been used in our experiments. We should report, however, that our results have always been consistent, independently of whether realistic (Studies 2 and 3) or less realistic (Study 1) faces were used. Finally, FaceGen also provides high-level parameters to change the physical appearance of the faces according to age, gender and ethnicity. These can be useful for future research exploring the impact of age, gender or culture on the effects reported in the dissertation.

Figure A.3. Faces created in FaceGen and integrated with the virtual humans platform.

Blushing, Sweating, Tearing and Wrinkles

Aside from bodily, vocal or facial expression, several autonomically mediated signals accompany emotions such as changes in coloration that result in local blood flow (e.g., flushing, blushing, blanching and bulging of arteries), whereas others involve additional detectable changes such as piloerection, sweating (and accompanying odors), tearing and crying (Levenson, 2003). The virtual human platform simulates some of these autonomic signals (de Melo & Gratch, 2009b; de

Melo et al., 2010a): blushing, sweating and tearing. Blushing, aside from being associated with self-consciousness, can be accompanied by social anxiety, uneasiness, embarrassment, guilt or happiness (e.g., when someone receives an undeserved praise; Leary, Britt, Cutlip, & Templeton, 1992). Blushing manifests physiologically as a spontaneous reddening of the face, ears, neck and upper chest as the small blood vessels in the blush region dilate, increasing blood volume in the area (de Graaf, 2002). Sweating is primarily a means of thermoregulation but can also be caused by emotional stress. This latter form is referred to as *emotional sweating* and manifests physiologically in the palms of the hands, soles of the feet, axillae and head (Kuno, 1956; McGregor, 1952). This form of sweating may occur in situations where an individual is subjected to fearful situations or the scrutiny of others (e.g., talking in public or to a superior). Crying is usually associated with the experience of intense emotions in situations of personal suffering, separation, loss, failure, anger, guilt or joy (Miceli & Castelfranchi, 2003). Crying is usually manifested by tearing and a characteristic loud noise (usually concealed with age). Finally, wrinkles are also simulated in the virtual human platform (de Melo & Gratch, 2009b; de Melo et al., 2010a). Typical wrinkle patterns occur with certain facial displays of emotion in people (Ekman, 1999b) and, thus, constitute an important cue for the identification of the emotion being expressed.

Blushing is simulated by a special *shader* (Akenine-Moller, 2008), running in the Graphics Processing Unit (GPU), which selectively applies a color tint over certain vertices in the face (e.g. the vertices in the cheek). Figure A.4 shows the expression of guilt using blushing. Simulation of tearing (and sweating) consists of modeling the properties of water and its dynamics using, once again, specialized shaders. Figure A.5 shows the expression of sadness with tearing and the expression of fear with sweating in the forehead. Wrinkles are simulated using bump mapping with normal maps (Akenine-Moller, 2008). Specific normal maps were defined for surprise,

sadness and anger (Figure A.6). Finally, a study was conducted where participants classified which emotion was being expressed in pictures of virtual humans expressing surprise, sadness, anger, guilt and fear with and without the autonomic signals (de Melo & Gratch, 2009b; de Melo et al., 2010a). The results showed significant improvement in the classification rates when using blushing, sweating, tearing and wrinkles.

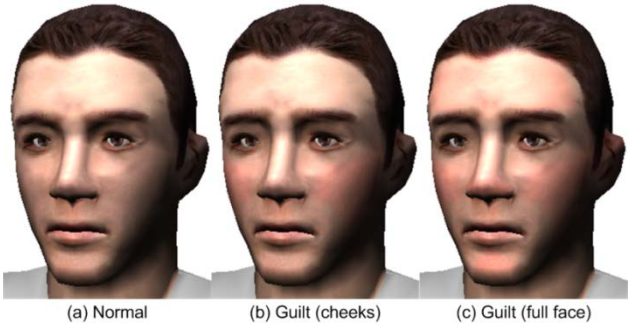


Figure A.4. Expression of guilt using blushing.

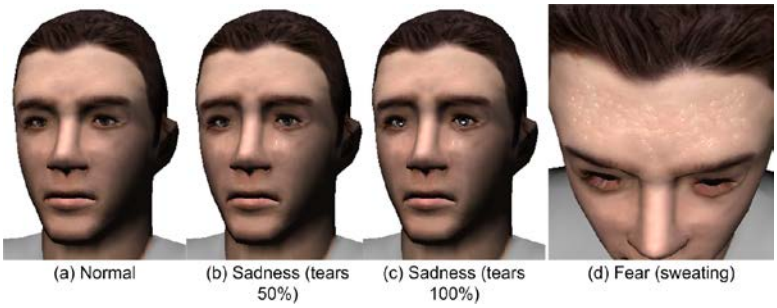


Figure A.5. Expression of sadness using tears and fear using sweating.

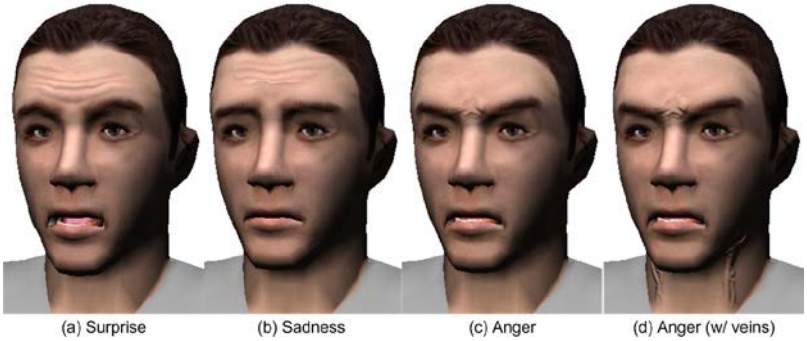


Figure A.6. Expression of surprise, sadness and anger using wrinkles.