UTRECHT UNIVERSITY

MASTER'S THESIS

FACULTY OF SCIENCE – ARTIFICIAL INTELLIGENCE

Expressing Emotions in the Iterated Prisoner's Dilemma

Author:
Sigurður Óli Árnason
Sn: 5961181

Supervisors:
Dr. Mehdi Dastani
Dr. John-Jules Meyer

20th of August 2019

Abstract

The mystery of cooperation has been researched from many perspectives and one explanation is that agents have fairness preferences where reciprocity matters. A key question in fairness literature is whether intentions matter and the modeling of fairness intentions turns out to require the tools of psychological game theory where payoffs depend not only on actions but also on beliefs. A problem in psychological game theory is that equilibrium analysis is difficult since beliefs are not typically observable ex post. I analyze how this affects behavior in the iterated prisoner´s dilemma, with agents that might care about fairness intentions, when one agent (Alice) cooperates and the other (Bob) defects. Alice thinks that Bob acted unfairly if he did so while believing that Alice would cooperate, but fairly if he did so while believing that she would also defect. In my analysis I define two types of players, a good one that cares about fairness intentions when choosing a strategy, and a bad one that only cares about himself and therefore always defects. Since Alice doesn't know ex post what Bob's belief was about her strategy, she has difficulties knowing his type and predicting his next move. This is a problem of asymmetric information and I look at the problem of signaling one's type to the other using signaling theory where the central problem is that for signals to be believable, they must be costly to fake. Emotions have been proposed as commitment mechanisms that help with this problem in human societies. I present a way to solve this by allowing players to express emotions between iterations of prisoner's dilemma. To do this I develop a theory of social affordance appraisal for agents, to argue that they can use emotional expressions as strategic actions in a social world.

Table of Contents

1 INTRODUCTION

1.1 MOTIVATION

Imagine that you betray someone after he trusted you not to. Humans have been shown to have strong feelings about fairness, justice, and reciprocity and such a situation often elicits emotions like guilt or shame in the transgressor. Caring about others' well-being like this seems to contradict the classic model of humans as rational self-interested beings but on closer look, it is not so simple. Intuition tells us that we would rather trust someone if we know that he will feel bad for betraying our trust, but how do we know that he feels bad? Previous attempts to model the effects that self-conscious emotions like guilt and shame have on behavior focus on how to include concerns about others' well-being when choosing actions and thereby deterring agents from acting selfishly. This is an *intra*personal approach but what is lacking in this research is the *inter*personal approach: What is the social function and implication of telling the others that he feels bad? In a perfect world with no perturbations these emotions will never be felt or expressed since the agent knows how to effectively choose an action that avoids that situation. The real world is not so perfect however, and an agent might choose an action that results in an outcome that he didn't intend. Expressing the right emotion in such situations can signal important information to others.

A classic tool to analyze other-regarding preferences is game theory and the prisoner's dilemma game, and fairness and reciprocity have been shown to be important concepts to reach an optimal outcome in this game. However, since intentions matter when assessing fairness of actions, modeling requires hierarchical beliefs about strategies since your belief about the other player's belief about your strategy is necessary to understand the intention of the action. When payoffs depend on such beliefs classic Neumann-Morgenstern game theory is not enough to model the game. Psychological game theory is a generalization of classic game theory that allows for analysis of such games but since beliefs, unlike actions, are not typically observable *ex post* (after the game ends) equilibrium analysis is difficult.

If we provide an opportunity for the agents to communicate their beliefs ex post, we encounter a signaling problem since agents with conflicting interests can't always be expected to send honest messages. For example, in the iterated prisoner's dilemma, if player 1 (Alice) cooperates while player 2 (Bob) defects, Bob has an incentive to make Alice think that it was not intentional, whether it was or not, since then Alice is more likely to forgive him and cooperate again in the next round. If she forgives him Bob gets a better outcome in the next round, whether he cooperates or not. The fundamental problem in such signaling games is that for a signal to be believable, it must be costly to fake, and emotions and emotional expressions have been proposed as a mechanism that humans use to send such costly-to-fake signals.

In the field of computational modeling of emotions, the typical approach is to view emotions as an *intra*personal force that affects decision-making. However, there is an emerging alternative approach in psychological literature of social-functionalism which views the emotions as *inter*personal forces that serve a role in social communication. This can be matched with the perspective of situated cognition on emotions where they are seen as actions in a social world that have a strategic social motive and function.

Guilt and shame are responses to one's own bad behavior and are two of the most researched moral emotions. Like for other emotions, the focus in the modeling of shame and guilt has been on the intrapersonal effects and the interpersonal role has been neglected. I will look at how agents can use emotional expressions of shame and guilt as signals that affect behavior in the iterated prisoner's dilemma by allowing them to use them in ex post signaling. This type of emotion signaling has not been modeled before in any way that I know. If I can model this, we can better see what we need to specify in a software agent so that agents in a multi-agent system or in communication with humans can use such mechanisms.

1.2 RESEARCH QUESTIONS

The research questions are then:

1) What is a shame signaling game?
2) What is a guilt signaling game?
3) Can signaling emotional expressions ex post help with cooperation in the iterated prisoner's dilemma?
4) Can software agents use emotional expressions to achieve and maintain cooperation?

1.3 OUTLINE

In the next chapter I present the prisoner's dilemma and the problem of ex post signaling from a game-theoretic perspective. In chapter 3 I present the social-functional perspective from psychology and the situated cognition perspective from cognitive science to argue that humans use emotional expressions as strategic actions in the social world, and then I introduce the emotions that I will use. In chapter 4, I explain how the agents can coordinate their strategies with the help of emotional expressions in ex post signaling. Chapter 5 reviews and discusses the results and chapter 6 presents the conclusions.

2 COOPERATION AND FAIRNESS

2.1 THE PRISONER'S DILEMMA

The problem of explaining how humans achieve and maintain cooperation is by now a classic. The difficulty is that mathematical models based on rational self-interested players contradict empirical evidence about human behavior (Fehr and Schmidt 2006). These models come from game theory and the model most used to analyze cooperation is the prisoner's dilemma game.

Game theory is a tool to study the strategic interaction between rational decision-makers in a formalized way and the foundation of the theory was laid in 1944 in the book 'Theory of Games and Economic Behavior' by mathematician John von Neumann and economist Oskar Morgenstern (Commemorative edition (Neumann and Morgenstern 2007)) (A more modern text on game theory is (Osborne and Rubinstein 1994)). The prisoner's dilemma is the standard example of a game analyzed in game theory and it demonstrates the problem of cooperation. In (Osborne and Rubinstein 1994) the game is described like this:

> *"Two suspects in a crime are put into separate cells. If they both confess, each will be sentenced to three years in prison. If only one of them confesses, he will be freed and used as a witness against the other, who will receive a sentence of four years. If neither confesses, they will both be convicted of a minor offense and spend one year in prison." (Osborne and Rubinstein 1994, p. 16)*

A general representation of the game can be seen in figure 1 where on top you see one suspects options and on the left you see the other one's options. The *cooperate* (C) option is to refuse to confess and stay true to one's partner while the *defect* (D) option is to sell him out. In the payoff cells the first number represents the payoff to the player on the left and the second number represents the payoff to the player on top. The payoff numbers are derived from a translation from years in prison to how good an outcome is, relative to the other outcomes, shifted so that all of them are positive. If the payoffs in figure 1 fulfill the condition $a > b > c > d$, then the game is a prisoner's dilemma. In the example above we see that $a = 4, b = 3, c = 1$, and $d = 0$ which fulfills the condition.

	Cooperate	Defect
Cooperate	b,b	d,a
Defect	a,d	c,c

Figure 1 - If a>b>c>d then this is a prisoner's dilemma game.

If both players have chosen a strategy and neither player gains from unilaterally changing his strategy after finding out what the other is going to do, then the set of strategy choices and the corresponding payoffs are called a *Nash equilibrium*. In the prisoner's dilemma the only Nash equilibrium is (D, D) but this is a suboptimal solution because both would be better off if they played (C, C). The suboptimal solution is a result of the incentive to defect; if a player is purely self-interested, then defecting is the rational strategy for both. The dilemma that the prisoners face is whether they should trust the other to cooperate to reach the optimal solution. If they have this trust between them, they can reach the optimal solution where neither confesses, otherwise they stay in the suboptimal solution where both confess.

In a one-shot game it is rational to defect, but if there is repeated interaction between agents and the players can learn from experience, then the game changes. In 1980, the political scientist Robert Axelrod set up a computer tournament where game theorists could submit strategies for playing an iterated version of the prisoner's dilemma and let those strategies compete against each other in a computer simulation (Axelrod 1980a, 1980b). The winning strategy of the tournament was called TIT-FOR-TAT which cooperates on the first move and then reciprocates what the other player does. In 1981, Axelrod and the evolutionary biologist William D. Hamilton published the landmark paper, 'The Evolution of Cooperation' (Axelrod and Hamilton 1981), that showed the importance of reciprocity in the evolution of cooperation.

2.2 FAIRNESS AND RECIPROCITY

(Fehr and Schmidt 2006) gives a good overview of evidence and theories of how people's concerns for altruism, fairness, and reciprocity motivate people. Up until the 1980's the belief that self-interest was the sole motivation for all people was widely held among economists. Over the last few decades, however, more research has focused on the topic and experimental evidence has shown that it is not the case that people are solely self-interested. It turns out that a substantial percentage of people are strongly motivated by other-regarding preferences. Fehr and Schmidt argue that the theory of self-interest is based on experiments and analysis in competitive markets where the achievement of other-regarding goals is impossible or infinitely costly. In environments outside of the competitive market, models that include other-regarding preferences do better in predicting people's behavior than the ones based on the self-interest assumption. A typical criticism of this approach to include other-regarding preferences in the model is that by changing preferences one is opening Pandora's box since everything can be explained by assuming the right preferences. They answer this criticism by saying that with the advancement in experimental tools, we now have the ability to gain insight into these preferences.

Fehr and Schmidt distinguish three main departures from the standard self-interest model made in the literature:

> *"In addition to the material resources allocated to him a person may also care about: (i) The material resources allocated to other agents in a relevant reference group. (ii) The fairness of the behavior of relevant reference agents. (iii) The "type" of the reference agents, i.e. whether the agents have selfish, altruistic, spiteful, or fair minded preferences." (Fehr and Schmidt 2006, p. 619)*

In the first case we have three types of behavior: Altruism, spitefulness (envy), and inequity aversion. *Altruism* is a form of unconditional kindness where the person gets happier as others get bigger payoffs irrespective of how it affects his own payoffs. *Spitefulness* is the opposite where one gets happier as others' payoffs are reduced, again, irrespective of how it affects his own payoffs. The third type is inequity aversion which is a conditional version of altruism and spitefulness. The conditional part is defined as the person's desire for equity. Therefore, an allocation that increases the equity of material payoffs increases his own payoff while an allocation that decreases the equity reduces his payoff. The outcome is of course dependent on the definition of equity which is very important.

The second case is often called *reciprocity*. A person responds to what he perceives as kindness with kindness and to what he perceives as hostility with hostility. This is based on the person's belief about the other one's intention. This type of analysis that looks into why others do what they do requires the tools of psychological game-theory (Geanakoplos et al. 1989) and Matthew Rabin modeled this in his pioneering article (Rabin 1993) for simple two-player normal form games. In the case of the prisoner's dilemma, it takes into account that if the other one defects, but you believe that he believed that you were also going to defect, then you think that was fair of him. The same goes for the case where you believe that he believes that you are going to cooperate that you think that it would be fair that he also cooperates.

In the third case it is the type of the agent that matters. One would behave kindly towards a "good" person and unkindly towards a "bad" person. The intention doesn't matter, just the perceived type. It can therefore be modeled using conventional game-theory.

(Falk et al. 2008) states that the most controversial question in the modeling of fairness preferences is whether the fairness of intentions plays a role or only the fairness of outcomes. The competition between those two theories has led to two different classes of fairness modeling: One that focuses on the fairness of outcomes based on inequity aversion, and the other that is based on the assumption that fairness intentions play a major behavioral role. The answer to this question is, according to them, of great practical and theoretical interest. They provide experimental evidence for the behavioral relevance of fairness attributions and conclude that models that combine intentions and outcomes fit their data best.

There are many ways in which the outcome can be incompatible with the intention, e.g. misunderstanding, miscommunication, temporary anger or frustration. A way to think of this is to imagine a perturbing function that the action intention goes through that might alter the intention into an outcome that is incompatible with the intention. We can see an interesting problem with the TIT-FOR-TAT strategy regarding this question whether an outcome is intentional. Imagine two players playing TIT-FOR-TAT that are in stable cooperation. If there is a perturbance and one player defects accidentally, they are thrown down an infinite alternating series of one player defecting while the other cooperates (see figure 2):

	Round 1	Round 2	Round 3	Round 4	…
TIT-FOR-TAT 1	C	C	D	C	…
TIT-FOR-TAT 2	C	D (accidental)	C	D	…

Figure 2 - Alternating defection in TIT-FOR-TAT vs. TIT-FOR-TAT.

The cooperation is therefore quite unstable if there is any noise, and to solve this a player needs a way to understand the difference between accidental and intentional defection. If it wasn’t intentional, then one can assume that he won’t defect again in the next round, so he should forgive the transgression and play cooperate.

2.3 RABIN FAIRNESS

(Rabin 1993) is one of the best-known economic models of fairness behavior where agents take intention into account in their fairness judgements. He bases his model on psychological game theory as introduced in (Geanakoplos et al. 1989) which is a modification of conventional game theory where one’s payoffs are not only based on actions but also on beliefs. By making the payoff a function of beliefs as well as strategies, he aims to incorporate these three facts about fairness behavior:

> *(A) People are willing to sacrifice their own material well-being to help those who are being kind.*
>
> *(B) People are willing to sacrifice their own material well-being to punish those who are being unkind.*
>
> *(C) Both motivations (A) and (B) have a greater effect on behavior as the material cost of sacrificing becomes smaller.*
>
> *(Rabin 1993, p. 1282)*

To do this, he sets up a model that defines players' strategies, belief about the other's strategy, and a second order belief about the other's belief about one's strategy. He then defines a kindness function that measures the intended kindness of actions based on the idea that if player 1 believes that player 2 is going to play a certain strategy, then, by choosing his strategy, player 1 is choosing the outcome that he believes will be the result from the interaction. If player 1 can choose from different strategies, he can choose different expected outcomes and those outcomes may have different material payoffs for player 2. If player 1 chooses the strategy that leads to the worst outcome for player 2 he is being unkind but if he chooses the strategy that leads to the best outcome for him, he is being kind. Using this kindness function, he defines a new utility function that incorporates facts (A) and (B) by making the players want to respond to kindness with kindness and unkindness with unkindness. To make his setup better for logical analysis I translate it to epistemic logic (for a background on epistemic logic see (Rendsvig and Symons 2019)).

The starting point for Rabin's analysis is a two-player normal form game with strategy sets S_1 and S_2 for players 1 and 2 where $\pi_i: S_1 \times S_2 \to \mathbb{R}$ is player i's material payoff independent of the kindness of the action. I describe the epistemic state of the agents before the game is played with a set of statements about the agents' chosen strategies and higher order beliefs. I use the epistemic operator Bel to signify beliefs and $Bel_i\varphi$ signifies that agent i believes statement φ. To say that player i is going to play strategy $\sigma\epsilon S_i$, I use the operator Act where $Act_i\sigma$ means that agent i has chosen to play strategy σ. $Bel_iAct_j\sigma$ then means that player i believes that player j is going to play strategy σ and $Bel_iBel_jAct_i\sigma$ means that player i believes that player j believes that player i is going to play strategy σ.

If player i's epistemic state is described by the statements $Act_i\alpha_i$, $Bel_iAct_j\beta_i$, and $Bel_iBel_jAct_i\gamma_i$ his epistemic state is defined by the values of α_i, β_i and γ_i where α_i is the strategy which player i intends to play, β_i is the strategy player i believes that player j intends to play, and γ_i is the strategy player i believes that player j believes that player i intends to play. For the two-player game we can present the epistemic state of the two players in a matrix, for a clearer presentation, which I call an *epistemic state matrix,*

$$\begin{bmatrix} Act_1\alpha_1 & Act_2\alpha_2 \\ Bel_1Act_2\beta_1 & Bel_2Act_1\beta_2 \\ Bel_1Bel_2Act_1\gamma_1 & Bel_2Bel_1Act_2\gamma_2 \end{bmatrix}$$

where $\alpha_1, \beta_2, \gamma_1 \epsilon S_1$ and $\alpha_2, \beta_1, \gamma_2 \epsilon S_2$.

To measure how kind player i is being to player j in this two-player game, Rabin defines the kindness function $f_i(\alpha_i, \beta_i)$ that takes two strategies as arguments and calculates how kind player i is by playing strategy α_i if he believes that player j is going to play strategy β_i.

The kindness function is supposed to answer the question of how kind player i is to player j by choosing to play α_i given his belief that player j intends to play β_i, since he is then choosing what outcome he intends to be the result.

Formally, if player i believes that player j is playing strategy β_i then by choosing to play strategy α_i, the material payoff pair he intends to be the result, for the two agents, is

$$\left(\pi_i(\alpha_i, \beta_i), \pi_j(\beta_i, \alpha_i)\right)$$

where $\pi_i(\alpha_i, \beta_i)$ is player i's material payoff independent of kindness if he plays α_i and player j plays β_i, and $\pi_j(\beta_i, \alpha_i)$ is player j's material payoff independent of kindness if he plays β_i and player i plays α_i.

He chooses this result from the set of all payoffs feasible if player j is choosing β_i, which is the set

$$\Pi(\beta_i) \overset{\text{def}}{=} \left\{\left(\pi_i(\sigma, \beta_i), \pi_j(\beta_i, \sigma)\right) | \sigma \epsilon S_i\right\}$$

Let $\pi_j^h(\beta_i)$ be player j's highest payoff in $\Pi(\beta_i)$ and $\pi_j^l(\beta_i)$ be player j's lowest payoff among the points that are Pareto-efficient in $\Pi(\beta_i)$. Let the "equitable payoff" be

$$\pi_j^e(\beta_i) = \left(\pi_j^h(\beta_i) + \pi_j^l(\beta_i)\right)/2$$

Finally, let $\pi_j^{min}(\beta_i)$ be the worst possible payoff for player j in $\Pi(\beta_i)$. Note that the only difference between this and $\pi_j^l(\beta_i)$ is that this includes points that are not Pareto-efficient in $\Pi(\beta_i)$.

Now we define the kindness function that measures how kind player i is being to player j as

$$f_i(\alpha_i, \beta_i) \overset{\text{def}}{=} \frac{\pi_j(\beta_i, \alpha_i) - \pi_j^e(\beta_i)}{\pi_j^h(\beta_i) - \pi_j^{min}(\beta_i)}$$

where $f_i(\alpha_i, \beta_i) = 0$ if $\pi_j^h(\beta_i) - \pi_j^{min}(\beta_i) = 0$.

In this definition of kindness, the part above the line is negative if player i chooses a payoff for player j that is lower than what is defined as equitable, and positive if he chooses a payoff that is higher than what is defined as equitable. Choosing the equitable payoff is therefore neither kind nor unkind. The part under the line is to normalize the function.

When player i is choosing whether he wants to treat player j kindly or unkindly he first wants to know if player j is planning on treating him kindly or unkindly, that is, he wants to know $f_j(\alpha_j, \beta_j)$. This requires the arguments α_j and β_j which player i does not know. However, by definition, β_i is player i's belief about

α_j and γ_i is player i's belief about β_j so we can replace α_j with β_i and β_j with γ_i and get player i's belief about how kindly player j is intending to treat him by calculating $f_j(\beta_i, \gamma_i)$. For clarity in future discussion we write $\tilde{f}_j$ to refer to player i's belief about player j's kindness and get

$$\tilde{f}_j(\beta_i, \gamma_i) = \frac{\pi_i(\gamma_i, \beta_i) - \pi_i^e(\gamma_i)}{\pi_i^h(\gamma_i) - \pi_i^{min}(\gamma_i)}$$

We are now ready to extend the material payoff function $\pi_i: S_1 \times S_2 \to \mathbb{R}$ with fairness preferences by using the kindness formulas. The new utility function defines the utility based on the player's beliefs and is defined by Rabin as

$$U_i(\alpha_i, \beta_i, \gamma_i) \overset{\text{def}}{=} \pi_i(\alpha_i, \beta_i) + \tilde{f}_j(\beta_i, \gamma_i)\big(1 + f_i(\alpha_i, \beta_i)\big)$$

This reflects facts (A) to (C) about fairness behavior that were previously mentioned. Let us assume fact (C) that says that the material cost of sacrificing is small, that is, that the changes in the second term of the formula for $U_i(\cdot)$ outweigh the first term. If player i believes that player j is being knowingly unkind, that is $\tilde{f}_j(\cdot) < 0$, then he will choose the α_i that treats player j unkindly to make $f_i(\cdot) < 0$ and maximize the utility function. On the other hand, if player i believes that player j is being knowingly kind, that is $\tilde{f}_j(\cdot) > 0$, then he will choose the α_i that treats player j kindly to make $f_i(\cdot) > 0$, to maximize the utility function. The kindness functions are bounded above and below so it follows that as the material payoffs get bigger, fairness concerns get less important. The relative power of material interests versus fairness concerns is something that needs a case-by-case analysis and is presumably highly dependent on the situation. However, the model can be used to get qualitative results.

Rabin uses the concept of psychological Nash equilibrium as defined by (Geanakoplos et al. 1989) to analyze the game, which he describes as such: "this is simply the analog of Nash equilibrium for psychological games, imposing the additional condition that all higher-order beliefs match actual behavior" (Rabin 1993, pp. 1287–1288). That is, both players are maximizing their utilities given their beliefs, and all beliefs are right so that

$$\alpha_i = \beta_j = \gamma_i$$

$$where \;\; i, j \in \{1,2\} \; and \; i \neq j$$

The intuition behind the condition that beliefs must match actual behavior is that when the game ends the outcome should be what they expected.

When the utility formula includes fairness concerns like $U_i(\cdot)$ does, he calls the psychological Nash equilibrium a *fairness equilibrium*.

2.4 FAIRNESS EQUILIBRIA IN THE PRISONER'S DILEMMA

Rabin applies this new model to a few different games, one of them being the prisoner's dilemma. To follow his analysis, we use the setup of the material payoffs as seen in figure 3.

	Cooperate	Defect
Cooperate	4X, 4X	0, 6X
Defect	6X, 0	X, X

Figure 3 - A scalable prisoner's dilemma.

According to Rabin, it turns out that we have a fairness equilibrium in this game where both cooperate if the material gains from defecting are small enough. This is true when $X < \frac{1}{4}$. This is because if it is common knowledge that they are playing (C, C) then both believe that the other is sacrificing their own material payoff to help the other (since defecting on the cooperating player would yield a higher material payoff). In this case neither wants to defect since that would give less utility from $U_i(\cdot)$ given the belief that the other is treating you kindly. That is, if the beliefs are such that $\tilde{f}_1(\cdot) > 0$ and $\tilde{f}_2(\cdot) > 0$, then they both choose the strategy which makes $f_i(\cdot) > 0$ to maximize utility. The cooperating fairness equilibrium is described with the following epistemic state matrix

$$\begin{bmatrix} Act_1C & Act_2C \\ Bel_1Act_2C & Bel_2Act_1C \\ Bel_1Bel_2Act_1C & Bel_2Bel_1Act_2C \end{bmatrix}$$

Mutual defection is also a fairness equilibrium and is independent of the value of X. In this case the common knowledge is that they are playing (D, D) and both want to satisfy their desire to hurt the other for not being willing to play cooperate and sacrifice X to give the other 6X. That is, if the beliefs are such that $\tilde{f}_1(\cdot) < 0$ and $\tilde{f}_2(\cdot) < 0$, then they choose the strategies that make $f_1(\cdot) < 0$ and $f_2(\cdot) < 0$ to maximize utility. The defecting fairness equilibrium is described with the following epistemic state matrix

$$\begin{bmatrix} Act_1D & Act_2D \\ Bel_1Act_2D & Bel_2Act_1D \\ Bel_1Bel_2Act_1D & Bel_2Bel_1Act_2D \end{bmatrix}$$

Equilibrium analysis is a big part of game theory, but it turns out that it is much more difficult in psychological game theory than classic Neumann-Morgenstern game theory. (Battigalli and Dufwenberg

2009) analyzes dynamic psychological games but they say that there is a reason to feel skeptical about a fully-fledged equilibrium analysis. They argue that players reach Nash equilibrium in conventional game theory by eventually having correct beliefs about the actions of the opponents after recurrent strategic interaction, but when payoffs depend on hierarchical beliefs, players need to learn those beliefs of others as well as their actions. The problem then being that while actions are observable ex post, after the game is finished, the beliefs are typically not. If, however, it is common knowledge that players make choices to maximize $U_i(\cdot)$ then one can deduce the beliefs ex post from the observed action.

For simplification I assume the internal consistency for each player that either $Bel_i Act_j C \wedge Bel_i Bel_j Act_i C$ or $Bel_i Act_j D \wedge Bel_i Bel_j Act_i D$ because otherwise we contradict the common knowledge that the other is maximizing $U_j(\cdot)$. I make this assumption based on the following reasoning. If $Bel_i Act_j C \wedge Bel_i Bel_j Act_i D$ we can write this as $Bel_i(Act_j C \wedge Bel_j Act_i D)$ but player i knows that it is irrational for player j to choose $Act_j C$ when his belief is $Bel_j Act_i D$ because choosing $Act_j D$ gives a higher outcome from $U_j(\cdot)$ (see appendix for proof). Therefore, if player i's beliefs are $Bel_i Act_j C \wedge Bel_i Bel_j Act_i D$ he contradicts the common knowledge that player j is maximizing $U_j(\cdot)$. By similar logic we can show the same for believing $Bel_i Act_j D \wedge Bel_i Bel_j Act_i C$.

In this setup of the prisoner's dilemma it turns out that $U_i(\cdot)$ is independent of the second order belief (see appendix for proof) and we could technically leave it out in further analysis. However, I keep it in for the sake of completeness.

2.5 GOOD AND BAD TYPES

It is a simplification to assume that both players make their choices based on $U_i(\cdot)$, and that it is common knowledge that they do. To make the analysis one step closer to reality I want to see what changes when there is a possibility that a player doesn't care about fairness intentions. Let us define a good and a bad type of players such that if player i is a good type he chooses the strategy that maximizes the utility function $U_i(\cdot)$ but if he is a bad type he only thinks about material payoffs and therefore tries to maximize $\pi_i(\cdot)$.

If player i is a bad type, he always chooses to defect because that gives the highest material payoff independent of the other player´s strategy. Assuming that $X < \frac{1}{4}$, if he is a good type he chooses to defect except if $Bel_i Act_j C$ and $Bel_i Bel_j Act_i C$ because then $\tilde{f}_j(\cdot) > 0$ and to maximize $U_i(\cdot)$ he maximizes $f_i(\cdot)$ which he does by choosing $Act_i C$.

Now it is more difficult to deduce the beliefs from observing the action because it is not known whether a player is maximizing $U_i(\cdot)$ or if he is maximizing $\pi_i(\cdot)$. To see the difficulty, consider the case where player 1 (Alice) cooperates but player 2 (Bob) defects, that is Act_1C and Act_2D. The only time a player i cooperates is when he is a good type and $Bel_iAct_jC \wedge Bel_iBel_jAct_iC$ so it must be that $Bel_1Act_2C \wedge Bel_1Bel_2Act_1C$. However, Alice cannot deduce Bob's beliefs because she doesn't know whether he is a bad type, in which case his beliefs could be anything, or a good type and $Bel_2Act_1D \wedge Bel_2Bel_1Act_2D$.

2.6 UPDATING BELIEFS

In Axelrod's tournament mentioned earlier, different strategies were designed that specify how to strategically choose an action based on previous outcomes when playing the iterated prisoner´s dilemma. When we extend the game to the psychological version, where the payoffs are not only based on the chosen strategies but also on beliefs, the players also need to choose what to believe before committing to a strategy for the next round. That is, we need to specify how players update beliefs between rounds. The strategy which the player then chooses to play in the next round is the one that maximizes the utility function, given his updated beliefs.

Consider the case where we have player 1 (Alice) and player 2 (Bob) that are good types and it is common knowledge that they are. In fairness equilibrium they either have stable defection or stable cooperation but if one defects and the other cooperates, they have a problem. Assume that Alice cooperates while Bob defects. Remember that in this case, where it is common knowledge that both players are good types, player i only cooperates if Bel_iAct_jC and $Bel_iBel_jAct_iC$ but if he defects then it must be that Bel_iAct_jD and $Bel_iBel_jAct_iD$. This means that the epistemic state matrix looks like this right before the round starts

$$\begin{bmatrix} Act_1C & Act_2D \\ Bel_1Act_2C & Bel_2Act_1D \\ Bel_1Bel_2Act_1C & Bel_2Bel_1Act_2D \end{bmatrix}$$

Now let us look at how to update beliefs after the round. The most straightforward belief-update strategy is to update the beliefs to what one assumes would have been correct for the round they just played. For Alice this means that she updates her beliefs to Bel_1Act_2D and $Bel_1Bel_2Act_1D$ while for Bob it would be to update to Bel_2Act_1C and $Bel_2Bel_1Act_2C$. Alice then chooses Act_1D to maximize her utility while Bob chooses Act_2C to maximize his utility. We then get the epistemic state matrix

$$\begin{bmatrix} Act_1D & Act_2C \\ Bel_1Act_2D & Bel_2Act_1C \\ Bel_1Bel_2Act_1D & Bel_2Bel_1Act_2C \end{bmatrix}$$

This results in a similar problem as when two TIT-FOR-TAT strategies are playing against each other and one defects accidentally causing an alternating series of one-sided cooperation.

A player could also try to anticipate how the other player is going to update his beliefs for the next round and then update his own beliefs accordingly. Let us assume that Alice does this and predicts that Bob is going to update his beliefs to Bel_2Act_1C and $Bel_2Bel_1Act_2C$ which will lead him to choose Act_2C since he is a good type. She will then keep her beliefs as they were, and we get the epistemic state matrix

$$\begin{bmatrix} Act_1C & Act_2C \\ Bel_1Act_2C & Bel_2Act_1C \\ Bel_1Bel_2Act_1C & Bel_2Bel_1Act_2C \end{bmatrix}$$

However, if Bob is also anticipating Alice to change her beliefs after observing the result of the last round, he expects her to change her beliefs to Bel_1Act_2D and $Bel_1Bel_2Act_1D$ which will lead her to choose Act_1D. He will then keep his beliefs as they were, and we get the unchanged epistemic state

$$\begin{bmatrix} Act_1C & Act_2D \\ Bel_1Act_2C & Bel_2Act_1D \\ Bel_1Bel_2Act_1C & Bel_2Bel_1Act_2D \end{bmatrix}$$

This is a communication problem and if there is an opportunity to communicate between rounds, a player could send a message to the other to tell him how he is going to update his beliefs and therefore what he intends to do in the next round. When it is common knowledge that both players are good types and that mutual cooperation is the best outcome for both players, if they message each other that they are going to cooperate in the next round, neither one has an incentive to break that promise and they therefore have a way to reach the cooperative equilibrium.

Such simple messaging is sometimes called cheap talk and (Farrell and Rabin 1996) explain concepts to analyze when cheap talk should be believed in classic game theory. It is not obvious how to apply their reasoning to a psychological game or that the logic is still valid, so I don’t apply it here. This is something that could be tackled in future research but I stick with the intuitive understanding that when an outcome is the one that has the highest payoff for both players and it is common knowledge that it is, then neither player has an incentive to lie about intending to play the strategy that would lead to that outcome.

Let us consider the case where there is a possibility that a player is a bad type. Bob knows that Alice is a good type since bad types always defect but Alice, on the other hand, cannot deduce Bob’s type from observing his action. This means that if Alice messages that she is going to keep her beliefs unchanged to give Bob another chance, Bob might want to keep defecting. Because of this doubt, they cannot solve the communication problem with simple messages because Bob might be a bad type that wants to defect on a player whom he expects to cooperate.

I want to address this problem of how the players can signal each other and coordinate beliefs and intentions between rounds when this doubt arises. This is a problem of asymmetric information, since a player knows his own beliefs and type, but the receiver of the signal doesn't, and we can analyze the problem with signaling theory.

2.7 SIGNALING

Signaling theory is a body of theoretical work on situations of information asymmetry between agents which tries to answer the question of when agents with conflicting interests should be expected to send honest signals and when they should be expected to cheat. It has its early roots in Michael Spence's (Spence 1978) analysis of signaling in job markets, but (Connelly et al. 2011) provides a concise synthesis of the theory and explains its key concepts. A key problem in such signaling games is that the signals must be hard or costly to fake. According to (Connelly et al. 2011) the cost of signaling is so central to signaling theory that some refer to it as the theory of costly signaling.

In game theory, interactions between agents where they take an action after each other are called extended games, and when a player only has partial information about the actions taken previously, they are called extended games with incomplete information. One such game is the *signaling game* which is at the heart of signaling theory. This is a Bayesian extended game, meaning that the players have beliefs with a known probability distribution about the incomplete information. (Osborne and Rubinstein 1994) describes a signaling game like this:

> *"There are two players, a "sender" and a "receiver". The sender is informed of the value of an uncertain parameter θ_1 and then chooses an action m (referred to as a message, though it may be payoff-relevant). The receiver observes the message (but not the value of θ_1) and takes an action a. Each player's payoff depends upon the value of θ_1, the message m sent by the sender, and the action a taken by the receiver." (Osborne and Rubinstein 1994, p. 237)*

If two players play a round of prisoner's dilemma where one cooperates and the other defects, the defector might for example want to send a signal to say that he defected because he thought the other was going to defect as well. He would like the cooperator to believe this and forgive him so that they can try to cooperate successfully in the next round. The defector is then a *sender*, the cooperator is a *receiver*, θ_1 states whether the sender is a good type that defected as a result of bad beliefs or whether he is a bad type, and the action is the signal that should convey the information from the sender to the receiver. We see the signaling dilemma here because if the sender is a bad type, he would also like the receiver to believe that he is a good type and forgive him so that he can take advantage of him once again. This means that the signal needs to be costly to fake to be believable.

2.8 THE COMMITMENT PROBLEM

Robert Frank’s book ‘Passions Within Reason: The Strategic Role of the Emotions’ (Frank 1988) makes the case that our emotions serve as biological *commitment mechanisms* that provide us with a way to effectively send such costly-to-fake signals about our mental attitudes. For example, when a person is trying to convince another person that he is honest, they can rely on the fact that we have biological features that make it hard for us to look like honest people when we are not. Even though this cannot be relied on perfectly, since there are people who can fake the signal, it makes it rational from a Bayesian perspective to trust the signal.

Another commitment mechanism mentioned by Frank is the ability to get angry and act temporarily irrationally. (Joffily et al. 2014) for example, shows that emotional reactions to free riders can lead people to apply costly sanctions on them in a way that doesn’t pay off on the short term. This deters others from taking advantage of you and treating you unfairly, leading to a better long-term payoff. It is hard to fake looking angry so if someone looks angry you assume that he is ready to sacrifice short-term payoffs to harm you.

The question, now, is if and how players can use emotions as signaling mechanisms to help coordinate their beliefs in the presented problem of iterated prisoner’s dilemma where agents care about fairness intentions.

3 USING EMOTIONAL EXPRESSIONS

3.1 AGENTS WITH EMOTIONS

After Herbert Simon introduced the theory of bounded rationality (Simon 1967, 1972) decision theory was forced to recognize that models of rational choice needed to account for situational and cognitive constraints. (Lerner et al. 2015) accounts how emotions have been shown to be an integral factor in human decision making. When Rosalind Picard started working on what would in 1997 become her seminal book *Affective Computing* (Picard 2000) her ideas of mixing emotion and computer science were met with laughter and she feared for her career at MIT (Picard 2010). Since then, much progress has been made and emotions are now accepted as a topic worth serious scientific studying. Computational modeling of emotions is becoming increasingly popular (See e.g. (Reisenzein et al. 2013) for overview) and one such technique is by implementing emotional BDI agents.

A BDI agent is a certain type of rational agent that has certain mental attitudes of *Beliefs*, *Desires*, and *Intentions* (Bratman 1987; Cohen and Levesque 1990; Rao and Georgeff 1991) and there have been attempts to extend BDI logic and architectures to include emotions (Meyer 2006; Pereira et al. 2007; Dastani et al.; Steunebrink et al. 2007). The psychological basis for this research is mostly the appraisal theory of emotion (Lazarus 1991; Frijda 1986; Ortony et al. 1990) which says that emotions are elicited by evaluations (appraisals) of events and situations. The emotion then influences the agent's motivation and may lead him to deal with the emotion with a coping strategy, specific to the emotion. BDI theory is apt for the study of emotional communication because it provides a way to reason about mental attitudes.

Appraisal theory tries to define the eliciting condition for each emotion and that condition defines a core relational theme. A core relational theme is a distinct theme that results in a distinct emotion and the eliciting conditions are measured with predefined appraisal dimensions. (Lazarus 1991) divides the appraisal into *primary appraisal*, which concerns whether and how the event or situation is relevant to the person's well-being, and *secondary appraisal*, which concerns the person's resources and options for coping with the event or situation. By calling it primary and secondary appraisal, Lazarus is not arguing for a sequential process. Rather it means that the primary appraisal is primary because it is responsible for the emotional "heat" of the emotion and a prerequisite for a strong emotion since if the event has no relevance to the agent's well-being then there is no reason for an emotion and the secondary appraisal is unimportant. The secondary appraisal is secondary because it is dependent on the primary appraisal.

He breaks secondary appraisal down to four components: Accountability (who gets the credit/blame for the harm/benefit), problem-focused coping potential (one's ability to act directly upon the situation), emotion-focused coping potential (perceived prospects of adjusting psychologically to the situation), and future expectancy (any reason independent of the individual's role for changes in the situation).

The idea of coping comes from the perspective that emotional behavior is a response to the emotion, originally seen as a way to deal with stress (Lazarus 1966; Pearlin and Schooler 1978). In this perspective, emotional behavior is not a part of the emotion, only a way to deal with it, to cope with it. The literature on emotional BDI agents, like most research on emotion, generally focuses on emotions as an internal state, an intrapersonal force, that influences the agent's evaluations and motivations, but there is room in these theories for a more social perspective on their effects and existence.

I present here, and argue for, an extended, interpersonal view that takes into account the empirical fact that our emotions not only affect our own behavior, but by expressing them to others they affect their behavior as well. If an agent knows that he can affect others' behavior with his emotional expressions, he has an incentive to use it strategically to his advantage to affect the world around him. To reason about how to use an emotion on an interpersonal level, I will define the concept of emotional expressions as afforded actions on others as a coping strategy.

3.2 THE SOCIAL-FUNCTIONAL PERSPECTIVE

> *"People use others' emotional expressions to infer traits and dispositions that are relevant to (social) survival and success (e.g., dominance, affiliation) and to anticipate others' behavior (e.g., collaboration vs. exploitation) as well as the trajectory of social interactions (e.g., cooperative vs. competitive)." (van Kleef 2016, p. 30)*

The strong focus on the *intra*personal effects of emotions in emotion research has been criticized and a different perspective that focuses on the *inter*personal effects has emerged (Parkinson 1996; Keltner and Haidt 1999; Keltner et al. 2006; van Kleef 2009; Fischer et al. 2008; Frijda and Mesquita 1994; Krueger 2012). This perspective is based on social-functionalism that assumes that emotions are best understood in the context of the social functions they serve and is inspired by the evolutionary approach to emotions proposed by Charles Darwin in his book *The Expressions of the Emotions* (Darwin and Prodger 1998), originally published in 1872.

While a more typical perspective on the functions of an emotion is about how it affects one's own behavior (intrapersonal), the social-functional perspective of an emotion is based on the idea that we are able to affect others' behavior by expressing our emotions (interpersonal). (Frijda 1986) already introduced the idea that others' behavior could be influenced by one's emotions and early empirical research to support this idea is a study on the effects of a mother's emotional expressions on the behavior of her child (Klinnert et al. 1983). The body of research on the interpersonal dynamics of emotion is now big enough that a whole book has been written on the topic (van Kleef 2016).

3.3 THE SITUATED COGNITION PERSPECTIVE

3.3.1 Situated emotions

To think of the emotions as more than internal states I resolve to the perspective of situated cognition. In cognitive science, the situated perspective on cognition is a challenge to the idea that cognition is something that only takes place inside the mind and argues that since minds are situated in a complex environment, a full study of the mind needs to take into account it´s interaction with it (see (Robbins and Aydede 2008)). (Griffiths and Scarantino 2005) apply this critique to emotion theory and propose a situationist's perspective.

A fundamental empirical support for the situated perspective on emotion is the study of audience effects where people are shown to respond differently to a constant stimulus depending on the expected recipients of the emotion. For example, when bowlers hit a strike they only smile once they have turned towards the audience (Kraut and Johnston 1979).

3.3.2 Affordance

When one expresses an emotion with a strategic social motive, it depends partly on the observer and the situation if the intended result is achieved. The term *affordance* is useful to analyze such situations. The psychologist James J. Gibson introduced the idea of (and the word) affordance as that which the environment offers the animal (Gibson 1977). It is not just an abstract property of an object, though, but something that is relative to the animal. In this way a chair affords sitting if it fits the size and posture of the animal in question. People and animals also provide affordances to other people, but these don't follow the laws of mechanics in the same way as other objects. "When touched they touch back, when struck they strike back; in short, they *interact* with the observer and with one another. Behavior affords behavior, and the whole subject matter of psychology and of the social sciences can be thought of as an elaboration of this basic fact." (Gibson 1977, p. 8).

Affordance theory has been used in the design of robotic agents (see (Horton et al. 2012) for overview) and one such approach is (Raubal and Moratz 2008) which provides an extended theory of affordances within a functional model for affordance-based agents. This extension goes beyond Gibson's ecological psychology of advocating knowing as a direct process, and supplements it with elements of cognition, situational aspects, and social constraints. They suggest that affordances belong to three different realms: physical, social-institutional, and mental. Physical affordance is the most basic concept of affordance which is defined by the physical properties of the environment relative to the person, such as the 'sittability' of a chair. Social-institutional affordances are the result of the imposition of social and institutional constraints of the environment on physical affordances. For example, a highway may physically afford driving at 200 km/h, but the institutional constraint of the law only affords driving as fast as the speed limit. Given such affordances to choose from after considering both physical and social-institutional constraints, the mental affordance is to decide which of the perceived affordances to utilize. The agent's behavior is therefore constrained by physical limitations, social-institutional limitations and mental limitations.

3.3.3 Social affordance appraisal

> *"Standard presentations of appraisal make emotions appear primarily perceptual. An emotion is the recognition that the world is a certain way. Representing the content of appraisals as affordances brings out the action-oriented nature of emotion." (Griffiths 2004, pp. 97–98)*

(Wilutzky 2015) explores the idea that emotions can resemble *pragmatic actions* since they are often aimed at achieving certain goals in a social context, and at other times they can be seen as *epistemic actions* when they perform the function of probing the social environment to extract or uncover important information. An emotional expression can in this way serve both as a way to signal one's position to others and to assess their intentions by reading into their response.

To draw out the (pragmatic) action-oriented nature of emotion (Griffiths 2004) presents the role of affordance in appraisal and calls it the 'Machiavellian emotion hypothesis' which is the idea that emotional appraisal is in part 'Machiavellian' or 'strategic'. He extends appraisal theory by introducing affordance as an appraisal dimension. By assuming that the expression of the emotion is an action on the world, this can be seen as a twist on Lazarus' appraisal dimension of problem-focused coping potential, the one that regards one's ability to act directly upon the situation. The evaluation of the affordances the social environment offers for coping potential can then be called *social affordance appraisal*.

3.3.4 Emotional expression as a coping strategy

As mentioned earlier, coping research originates in stress theory and is typically thought of as a way to deal with distress by changing one's own mental attitudes (e.g., beliefs, desires, or intentions) (see e.g. (Dastani and Lorini 2012)). This view is limited in its ability to explain emotional behavior in a social context from a social-functional perspective. For example, coping with anger by changing one's own mental attitude doesn't fully explain its social function. It is only by expressing it and changing others' mental attitudes that the emotion serves its social function of making others afraid to take advantage of him.

(Lazarus 1991) does not include the coping-potential components as important appraisal components when he defines the elicitation conditions for anger and guilt, and he defines low problem-focused coping ability as an important appraisal component for sadness. From the social-functional view, high social affordance appraisal should be a factor for all of them since they only serve the social function if they affect others. However, by seeing the expression of the emotion as a way in itself to cope with the emotion, the problem-focused coping potential dimension makes room for the social-functional perspective. By expressing an emotion one can then, instead of coping by changing one's own mental attitudes, try to change others' mental attitudes. In other words, if emotion expressions are seen as actions that can change the world, then the problem-focused coping potential appraisal can include social affordance appraisal and result in an emotional expression being a coping strategy.

3.4 THE MORAL EMOTIONS

The fundamental factors in primary appraisal, as presented by (Lazarus 1991), concern whether and how the event or situation is relevant to the person's well-being. However, as has been said, it has been shown that it is not the case that people are solely self-interested and it turns out that a substantial percentage of people are strongly motivated by other-regarding preferences.

Emotions that are linked to the interests of others are called the moral emotions. Jonathan Haidt (Haidt 2003) categorized them into four families: Other-condemning (anger, contempt, disgust), self-conscious (guilt, shame, embarrassment), other-suffering (compassion), and other-praising (elevation, gratitude). He argues for the same perspective as Rabin does, that morality and reciprocity only function as a two-edged sword since it is not enough to be kind to those who are kind, it is also necessary to punish or condemn those who are unkind. He proposes that the self-conscious emotions developed as an answer to others' other-condemning behavior.

Of the moral emotions, maybe the most focus in research has been on guilt and shame, and the difference between them has been found to be subtle but important. Shame is a feeling that the core self is defective and causes the person to withdraw from interaction with others. Guilt, on the other hand is focused on the action itself being bad and motivates the person to try and make up for the harm that was caused and repair the relationship that was harmed by the action (Haidt 2003; Tangney et al. 2007). I give here an account of the other-condemning family and the self-conscious family.

3.4.1 Other-condemning

Haidt argues that anger is, usually, unfairly thought of as an immoral emotion. He thinks that this is unfair because it is not just a primal urge for violence but also the emotion that makes people stand up and demand justice. He gathers from the literature that in people's descriptions of elicitation of anger "themes of frustration and goal blockage mixed with more moral concerns about being betrayed, insulted, and treated unfairly" (p. 856). It can be a reaction to such situations for either oneself or a friend. He gathers that "anger generally involves a motivation to attack, humiliate, or otherwise get back at the person who is perceived as acting unfairly or immorally" (p. 856).

Disgust, he gathers, is a distaste response that serves as a guardian of the temple of the body in a broad sense. Evolving from a more basic physical disgust often relating to the sense of taste and smell, to the more complex sociomoral disgust where it serves as the guardian of the lower boundary of the category of humanity. About the action tendencies he says: “All forms of disgust include a motivation to avoid, expel, or otherwise break off contact with the offending entity, often coupled to a motivation to wash, purify, or otherwise remove residues of any physical contact that was made with the entity” (p. 857). Sociomoral disgust results in ostracizing and is based on the person himself being condemned and not just the action.

Contempt falls between anger and disgust and is almost a mix of the two. It is usually thought of as being elicited from a feeling of being morally superior or looking down on the other. The data is scarce, though, and in experiments it is often confused with what is better understood as disgust. Contempt seems to be a cool emotion that doesn’t motivate action, but rather it changes the treatment that the object of the emotion gets in the future in a way that he receives less warmth, respect, and consideration in future interactions.

The CAD-triad hypothesis (Rozin et al. 1999) describes a well-documented relation between these emotions and Schweder’s three moral codes, the ethics of community, autonomy, and divinity. It says that contempt, anger, and disgust are responses to these moral codes, respectively, so that anger is linked to autonomy (individual rights violations), contempt to community (violation of communal codes, including hierarchy), and disgust to divinity (violations of purity-sanctity). In this view, the other-condemning emotions serve as guardians to different parts of the moral code. All of the three emotions motivate a change in one’s relationship with the moral violator, but anger is the only one that motivates direct prosocial action.

3.4.2 Self-conscious

Haidt argues that since we have a strong need to belong to groups we would inevitably adapt to others’ other-condemning behavior by regulating our own behavior, and the self-conscious emotions seem designed for this purpose. In western culture the principal emotions in this self-conscious emotion family are shame, embarrassment, and guilt. Pride can also be included as the positive opposite to shame. Most Asian cultures, however, do not lexically distinguish between shame and embarrassment, and in some non-Western cultures guilt does not even exist. Haidt makes sense of these differences from the fact that these emotions depend on the cultural variability in whether the self is seen as independent or interdependent, and whether the social structure is hierarchical or egalitarian.

Haidt summarizes the literature about *shame* in Western cultures as being “elicited by the appraisal that there is something wrong or defective with one's core self, generally due to a failure to measure up to standards of morality, aesthetics, or competence” (p. 860). *Embarrassment*, on the other hand, ”is said to be elicited by appraisals that one's social identity or persona within an interaction is damaged or threatened, most commonly because one has violated a social-conventional rule but also at times because of events beyond one's control” (p. 860). The action tendency for both is to “reduce their social presence, create a motivation to hide, withdraw, or disappear, and making movement and speech more difficult and less likely” (p. 860). For Westerners, the action tendencies are milder for embarrassment than shame.

While shame is linked to hierarchical interactions, guilt seems to grow out of communal relationships and the attachment system. It is not elicited from simply appraising that one caused harm, but from realizing that the harmful action threatens one’s relatedness to the victim. Guilt motivates people to help one’s victim or help make up for one’s transgression to restore or improve their relationships.

(Tangney et al. 2007) gives a detailed overview of the literature and research on the difference between shame and guilt and supports Haidt’s distinction between them, saying that guilt is directed towards one’s action being bad instead of the self in its entirety, as it is for shame. They claim that guilt goes hand in hand with other-oriented empathy while shame disrupts the ability to form empathetic connections with others. This matches the different action tendencies for the two emotions. Shame causes people to focus on how they exposed their flawed self and go into hiding whereas guilt causes people to look outwards on the harm they caused to others and try to make it better.

In her book, ‘The chrysanthemum and the sword: Patterns of Japanese culture’ (Benedict 2005), first published in 1946, the anthropologist Ruth Benedict described Japanese culture as shame culture and US culture as guilt culture, and since then research has shown significant cultural variation in valuation, elicitors and behavioral consequences of shame and guilt. (Wong and Tsai 2007) give an overview of these differences. They argue that the reason for the cultural variation is that different cultures view the self differently. Individualistic cultures view the self as independent and to be differentiated from one’s temporary actions while collectivistic cultures view the self as interdependent and so other people’s feelings are as important as one’s own. They review empirical evidence and argue that in collectivistic contexts there is less difference between shame and guilt than in individualistic cultures because then people don’t view themselves as separate from their relationships to others. They argue that while shame is seen as harmful to psychological well-being in individualistic societies, it is seen as healthy and necessary in collectivistic societies. (Tangney et al. 2007) reviews the evidence on harmfulness of guilt and shame and supports the statement that shame is psychologically harmful, and guilt is not, but the research has a strong Western bias.

3.4.3 Guilt-eliciting and shame-eliciting emotions

We see from the above that guilt and shame serve very different functions. Since Haidt argues that the self-conscious emotions developed as an answer to others’ other-condemning behavior I draw a conclusion, from the social-functional perspective, that shame and guilt are elicited after social affordance appraisals of different types of judgements. Since shame is a feeling of a flawed self, it is reasonable to assume that it is a response to another's judgement of his self. Guilt on the other hand is a feeling of one's action being bad, so it is reasonable to assume that it is a response to a judgement of one's action.

The other-condemning emotions Haidt mentions serve the function of condemnation, but guilt arises from empathy rather than a fear of condemnation. It seems that if one wants to elicit guilt in the other, there is a role to be found for the softer *supplication* emotions such as sadness and disappointment that signal need for help and elicit empathy in others.

Disappointment signals unfulfilled expectations (Lelieveld et al. 2012) and (van Kleef et al. 2006) explains that the dynamic between disappointment and guilt is helpful for cooperation. (Lelieveld et al. 2012) shows how, in bargaining situations, anger evoked a complementary emotion of fear if the condemnation came from a high-power bargainer but a reciprocal emotion of anger when it came from a low-power bargainer. When a bargainer expressed disappointment, it evoked a complementary emotion of guilt in the other, independent of power balance.

There are no perfectly clear distinctions between these emotions but to talk about which emotions elicit guilt and which emotions elicit shame, my conclusion is that there are generally two directions one can go. The one that elicits shame is *condemnation* which is a mix of anger, contempt, and disgust, that threatens the person socially or physically. The mix depends on which of the CAD-triad the condemner believes applies to the situation and his position to make threats. The risk of choosing this path is that threats can be met with responding threats, anger can be met with anger, which sends both players into a downward spiral. The second is the path of *supplication* which has the goal of eliciting guilt in the other by making it known to him that he hurt you with his action. Simple sadness is not always enough since guilt is often a feeling of not living up to expectation, and disappointment enforces that belief where sadness is less specific about the cause of the harm. The risk of choosing this path is that one can be ignored and left with the disappointment and no resolution or change. One can say that the social function of disappointment is to elicit guilt and the social function of the other-condemning emotions is to elicit shame. The next chapter looks at how we can use these two pairs of emotions, condemnation/shame and disappointment/guilt, to solve the problem in chapter 2.

4 EMOTIONAL COMMUNICATION

4.1 TYPE-SIGNALING

I return, now, to the problem of coordinating beliefs and strategies ex post in the iterated prisoner's dilemma. Consider the situation after player 1 (Alice) cooperates and player 2 (Bob) defects. Remember that the problem is that Alice doesn't know whether Bob is a good type that defected because he believed Alice was going to defect, or a bad type that defects independent of his beliefs. Bob doesn't know what she decides to believe so he doesn't know what to expect in the next round. In this example, Alice must be a good type since a bad type never cooperates. If they can achieve a state of common knowledge that both players are good types, then they can coordinate on the cooperating equilibrium by messaging each other, as explained before, because then they know that their interests are aligned.

Primary appraisal of the situation makes Alice emotional about Bob's defection, but secondary appraisal determines the exact emotional response. A part of the secondary appraisal is the social affordance appraisal that strategically chooses an emotion that has a chance of serving its social role, depending on the affordances of the situation. It seems like a paradox to say that an emotion is hard to fake but still strategically chosen, but my assumption is that it is the primary appraisal that is hard to fake while the secondary appraisal chooses from a set of possible emotions, given the primary appraisal. I have presented disappointment and condemnation as two possible paths which the secondary appraisal should consider. The social function of disappointment is to elicit guilt in the other player but if Alice does not believe that there is a chance that Bob is a good type then the situation does not afford disappointment because Bob does not care about her well-being and will not feel guilty. The social function of condemnation is to make threats and elicit shame but the situation might not afford making threats, for example if Bob is more powerful than Alice or, as in our example, there is simply no mechanism to punish the other that one can threaten to use.

If Alice expresses disappointment, then Bob's primary appraisal will only elicit an emotion if he is a good type because otherwise, he doesn't care about Alice's well-being. Since disappointment does not just signal that Alice feels bad about the outcome but also that she expected more from Bob, his social affordance appraisal will see an affordance for forgiveness. This elicits guilt in him and by expressing the emotion he sends a hard-to-fake signal that tells Alice that he is a good type and that he wants to make up for his transgression. We therefore have a litmus test that determines Bob's type. This effectively solves the information asymmetry of Bob's type and the players can reach the cooperative equilibrium through simple communication if it proves Bob to be a good type, but the defective one if it proves him to be a bad type. If Alice believes that there is a chance that Bob is a good type, she sees an affordance for disappointment but if she is sure that he is not, there is no point in expressing disappointment. We see that in this example the role of the emotions as signals is to reduce information asymmetry about players' types. Being of a certain type means that one operates according to a certain formula that is associated with being of that type. The belief coordination then happens through deliberation and simple messages that are believable only when it is known that the player is of a certain type.

If Alice chooses condemnation, she has decided that Bob is a bad type and does not invite him to prove otherwise. The social function of condemnation is based on threat-making, social or physical, but in this setup of the iterated prisoner's dilemma there are no opportunities to make threats. If such mechanisms were present, then the condemning emotions are signals in a complex threat-making game that one is ready to sacrifice one's own payoff to punish the other, while shame would be a signal of submission that says that he is going to stay away from future interactions to avoid punishments. This can be seen as type-signaling where the type describes one's power, or how much he cares about his reputation, instead of one's fairness concerns. To model this, we would need a different and more complex game which I will not do here. If there is no such mechanism available, Alice does not condemn Bob because there is no way to make any threats, social or otherwise. We could say that the situation affords neither condemnation nor shame. We also see that disappointment is a better candidate to reach the cooperating equilibrium since it aims at forgiveness while condemnation aims to drive the transgressor away from future interaction.

4.2 USE IN SOFTWARE AGENTS

To implement this in a software agent, one can try and implement the social affordance appraisal and include it in the agent's deliberation. For such emotion-communication mechanisms to work in a society of agents, there needs to be a high enough ratio of them that have the same emotion mechanisms and common knowledge about each-others' emotional workings. That is, the agents need to have a reason to believe that the other is a good type that is constrained by these emotional commitment mechanisms. I believe that an understanding of these mechanisms is an integral part of emotional intelligence and that the formal mapping of emotional communication is necessary to develop truly emotionally intelligent agents.

In research and development of agents and robots, the goal of autonomy is increasingly important, and robots are shifting from being tools that are operated by humans to being more sophisticated partners. As we have seen in this thesis, however, sometimes interdependence between agents is beneficial. (Johnson et al. 2010) present the idea of *coactive design* which builds on the concept of teamwork-centered autonomy and takes into account the fact that people often work together in complex ways to solve problems. They say that this interdependence of participants in joint activity is a critical factor in the design of human-agent systems. I believe that this interdependence can be implemented, at least in part, through emotions.

One of the biggest unanswered questions is how to implement the social affordance appraisal but answering this is beyond the scope of this thesis. If an agent is meant to communicate with emotional expressions in interactions with humans, this requires him to have complex and nuanced skills to understand and assess the situation. He must for example be aware of the cultural differences in the social functions of emotions like we see for example in the differences between guilt cultures and shame cultures.

Another problem is that it must somehow be reinforced that agents cannot fake certain emotions, or that it is costly for them to do so. This requires trust in the design of the agents which would have to be based on trust in the designer. We humans get this from our biological design but unless there is an agreement and trust between designers of robots to implement their agents with such limitations, this cannot be counted on for those agents. However, if there is trust that agents have these limitations, they gain other abilities as explained in this thesis. This would all be simpler for inter-agent interaction than human-agent interaction since they can then be implemented to have the same simplified model of how emotional communication works.

5 RESULTS AND DISCUSSION

I set out to find a solution to the problem of beliefs not being observable ex post in the iterated prisoner's dilemma by using guilt and shame as signaling mechanisms. Based on the idea of fairness intentions, I defined formally a specific case of two players playing the iterated prisoner's dilemma where the players are either good types that care about intentions so that they choose the same action they believe the other is going to choose, or bad types that don't care about intentions and always defect. I showed that this leads to an information asymmetry if one agent defects while the other one cooperates because the cooperating agent doesn't know whether the other one defected because he is a bad type or because he is a good type that thought that the other one was going to defect as well. To fix the information asymmetry with signaling, the signals need to be costly to fake, since there is possibly an incentive to lie, and I provided evidence from psychology based on the social-functional perspective to argue that emotional expressions are used as such costly-to-fake signals and that they can be used to reduce the information asymmetry. I do this within the framework of appraisal theory by including an appraisal dimension which I call social affordance appraisal where emotions are elicited after an evaluation of whether they can serve their social function in the current social setting. As appraisal theory has been widely used for computational modeling of emotions it was important to ground my arguments about the emotions in this theory so that my results can be used to extend existing models.

I argued that when an agent (Alice) feels that another agent (Bob) has treated her unfairly, she expresses an emotion that serves a social function by asking for an emotional response from Bob. The social function of Alice's emotional expression is to reveal information about Bob's type by looking at his emotional reaction. I looked at two paths which Alice can take when she feels that Bob has treated her unfairly. The first path is to express disappointment with the goal of revealing whether Bob is a good or a bad type by seeing if his reaction is to express guilt. Since only a good type responds to disappointment with guilt this serves as a litmus test that solves the information asymmetry about Bob's type. The second path is to condemn Bob by expressing any of the three other-condemning emotions (anger, contempt, or disgust) with the aim to elicit shame. My conclusion is that if Alice chooses this path then she has decided that Bob is a bad type and is not asking him to prove otherwise. Instead, she is attempting to elicit shame in him by signaling that she considers him to be a bad type and that he should withdraw from future interactions or risk punishment. In the setting I provide, there is no mechanism to withdraw from interaction nor to threaten punishment, so the situation does not afford condemnation nor shame.

When Alice runs the litmus test and reveals that Bob is a good type, both know that the optimal outcome for the other one in each iteration of the prisoner's dilemma is mutual cooperation. They can then message each other that they are going to cooperate in the next round, and neither one has an incentive to break that promise. If the litmus test reveals that Bob is a bad type, then both will defect in the next round independent of any messaging between them. This effectively reduces the information asymmetry and provides the good players with an ability to overcome perturbations in order to maintain stable cooperation.

This dynamic is an extremely simplified abstraction of what happens in real human situations. A more accurate model would include ways to analyze what happens when agents try to trick the other by faking signals and play power games using complex social structures. Modeling this behavior becomes exponentially more complex with every additional variability in behavior but I believe that further research would be beneficial both for developing emotionally intelligent agents and also as a paradigm in computational modeling of emotions for further psychological research.

My initial idea was that emotions would signal beliefs directly, so that other means of messaging were unnecessary, and the fairness equilibrium would be reached after a conversation of sorts where emotional expressions are the way of communicating. This turned out to be very complicated and there was not enough psychological evidence to argue for a system that could effectively solve the problem. I found more support in the literature for using these chosen emotions to discern types rather than beliefs, so I chose to take that path instead to solve the information asymmetry. This was a disappointment to me since much time and effort went into trying to set up a system that communicated beliefs directly. This required a complex mix of emotion theory, psychological game theory, signaling theory, and epistemic logic, which, I found, has not been developed far enough to be applied to solve a problem like the one I wanted to solve. A model that synthesizes these diverse theories and disciplines to better describe the logic behind the communication of beliefs and intentions through emotional expressions should be a very exciting topic for future research.

There is reason to believe that emotional expressions signal beliefs directly in many cases and it is worth continuing research in that direction. One new and interesting perspective that analyzes emotional expressions as communicative devices, but was not touched upon here, is Andrea Scarantino's Theory of Affective Pragmatics (Scarantino 2017) that compares them to *speech acts* a la Searle (Searle 1969) and Austin (Austin 1975) and is based on the insight that it is possible to engage in analogs of speech acts without using language at all. This seems like a good point to work from for future research in artificial intelligence that aims to develop a theory of multi-agent systems where the agents can use emotional expressions to communicate.

I would be remiss not to point out the ethical concerns of researching how to manipulate people using their emotions. Like in many areas of scientific research, the results I present can be used for both good and bad, and I see enormous benefits in agents knowing how to use emotions strategically. If we want to design agents with human-like intelligence, the danger follows that they can use those abilities for bad, just like humans do, but it also provides a way to design them in a way that they understand how to be good.

6 CONCLUSION

These were the research questions:

1) What is a shame signaling game?
2) What is a guilt signaling game?
3) Can signaling emotional expressions help with cooperation in the iterated prisoner's dilemma?
4) Can software agents use emotional expressions to achieve and maintain cooperation?

I found that the best way to look at emotion signaling is through the perspective of social-functionalism where the expression of an emotion is viewed as having a social function. Guilt and shame are responses to a situation where one feels he has harmed another agent, but they serve different social functions and are elicited after different appraisals of the situation. I made the case that the appraisal process, which is responsible for deciding what emotion is elicited, is partially strategic and will only elicit an emotion if the agent believes that the emotion has an opportunity to serve its social function, given the situation. I say that the situation affords the emotion if this is the case. Shame is a feeling that the core self is defective and causes the person to withdraw from interaction with others. Guilt, on the other hand is focused on the action itself being bad and motivates the person to try and make up for the harm that was caused and repair the relationship that was harmed by the action. What decides which one is elicited is the emotional response from the other agent which signals information about which strategy the situation affords.

In a situation where Bob just treated Alice unfairly, a guilt signaling game is a situation where Bob sends a hard-to-fake signal to Alice to prove to her that her well-being is a factor in his utility formula and that he didn't intend the unfair outcome. When it is common knowledge that agents include others' well-being in their utility formulas, it is easier for them to coordinate on desired outcomes using simple messages. A shame signaling game is a situation where Bob sends a hard-to-fake signal to Alice to convince her that he will stay away from future interaction in order to avoid punishment that has been threatened.

Bob sends these signals as a response to an emotional expression from Alice about the situation. If Alice expresses disappointment, which is a judgement of the action instead of Bob's character, Bob perceives affordance for forgiveness and therefore plays the guilt signaling game. If Alice expresses any of the three other-condemning emotions (anger, contempt, or disgust) he does not perceive affordance for guilt since these emotions are a judgement of character and not just the action. Instead, he sees a need to avoid the punishment that is an implicit threat that underlies the condemning emotions and therefore he signals shame to prove to Alice that he will stay away from future interactions. Both shame and guilt are therefore responses to Alice´s emotional expressions that have the purpose of signaling certain information about him.

Guilt signaling can help with both maintaining and achieving cooperation in the iterated prisoner's dilemma if there are some agents that care about fairness in a way that they prefer mutual cooperation over the outcome where they defect on a cooperating agent, and only those agents can signal guilt. If Alice cooperates while Bob defects, then Alice can express disappointment to see if Bob is a good or a bad type, and if he proves to her that he is a good type by signaling guilt they have enough trust between them to coordinate effectively on mutual cooperation in the next round by simply telling the other that they are going to cooperate. In a normal iterated prisoner's dilemma, there is no affordance for condemnation nor shame because there are no mechanisms to make threats and agents don't have a choice whether they want to withdraw from future interactions or not. Adding such options makes the game more complex and I did not do that in this thesis. This means that the analysis of the shame game was somewhat limited.

If there are enough agents that are constrained by fairness concerns, and they can signal these concerns successfully via emotional expressions, the agents have a way to overcome perturbations in an effective manner and maintain cooperation. A system of software agents can use this if there is trust that the signals are hard to fake, and it is common knowledge what the signals mean. The key principle is to make the primary appraisal involuntary while the secondary appraisal is strategic.

Further, endowing an agent with emotions that affect his rationality in the same way as humans, and are communicated via emotional expressions that are therefore understood by humans, can be a paradigm to implement interdependent agents for human-agent interaction. I argued that there is room in the appraisal theory of emotion to implement emotional expressions as strategic actions and this is a path to implementing it in software agents since appraisal theory has been widely used in computational modeling of emotion.

This way of explaining and analyzing the social functions of emotions opens up new possibilities in emotion modeling and development of emotional agents. These principles can be used as inspiration for agent-to-agent communication, design paradigms in human-robot interaction, or a guide for psychological research. It gives us a new and interesting theoretical perspective on the nature of emotions that has been neglected in many disciplines.

I left out any discussion about social and moral norms, but these have been proposed as factors when talking about guilt and shame. Research on norms focuses on how agents learn to adjust their utility formulas, but I was more interested in interpersonal dynamics if we take those as given. I see, however, many opportunities in future research to analyze norm enforcement and emergence from the perspective of emotion signaling. Emotional expressions could be used as messages between the agents as a form of reinforcement learning to inform each other about whether they think the behavior of another is acceptable or not, and how serious their offense is. This could include the other moral emotions such as pride, gratitude, and elevation for positive reinforcement, as well as the ones mentioned in this thesis which would be used for negative reinforcement. To draw out an even more social nature of emotional expressions, agents could observe the emotional expressions of agents judging a third party and learn from that judgement.

What I have presented is an attempt to show that researchers in computational modeling of emotion and design of emotional agents should expand their approach from a strictly intrapersonal perspective and look at the possibilities presented when we allow for the emotions to be interpersonal phenomena. I think that I have done so convincingly and shown that there is important unexplored territory to be mapped out in the world of emotional communication.

7 Appendix

Proposition:

In the prisoner's dilemma game described in the following table with $0 < X < \frac{1}{4}$, if player i is a good type, he will choose Act_iC if Bel_iAct_jC and he will choose Act_iD if Bel_iAct_jD independent of whether his second order belief is $Bel_iBel_jAct_iC$ or $Bel_iBel_jAct_iD$.

	Cooperate	Defect
Cooperate	4X, 4X	0, 6X
Defect	6X, 0	X, X

Figure 4 – A scalable prisoner's dilemma.

Proof:

This table describes the material payoffs, but we must calculate the possible outcomes for $U_i(\cdot)$. First, we calculate the possible outcomes for the kindness functions $f_i(\cdot)$ and $\tilde{f}_j(\cdot)$. Remember that $f_i(\cdot)$ and $\tilde{f}_j(\cdot)$ are formally equivalent, even though they were made notationally distinct, so they give the same outcome for the same input. Also note that the game is symmetric so any calculation for player i is the same for player j.

$$f_1(C,C) = \frac{\pi_2(C,C) - \pi_2^e(C)}{\pi_2^h(C) - \pi_2^{min}(C)} = \frac{4X - \frac{4X+0}{2}}{4X - 0} = 1/2$$

$$f_1(C,D) = \frac{\pi_2(D,C) - \pi_2^e(D)}{\pi_2^h(D) - \pi_2^{min}(D)} = \frac{6X - \frac{6X+X}{2}}{6X - X} = 1/2$$

$$f_1(D,C) = \frac{\pi_2(C,D) - \pi_2^e(C)}{\pi_2^h(C) - \pi_2^{min}(C)} = \frac{0 - \frac{4X+0}{2}}{4X - 0} = -1/2$$

$$f_1(D,D) = \frac{\pi_2(D,D) - \pi_2^e(D)}{\pi_2^h(D) - \pi_2^{min}(D)} = \frac{X - \frac{6X+X}{2}}{6X - X} = -1/2$$

We then get

$$U_1(C,C,C) = \pi_1(C,C) + \tilde{f}_2(C,C)\big(1 + \ f_1(C,C)\big) = \ 4X + \frac{1}{2} \cdot \left(1 + \frac{1}{2}\right) = 4X + \frac{3}{4}$$

$$U_1(D,C,C) = \pi_1(D,C) + \tilde{f}_2(C,C)\big(1 + \ f_1(D,C)\big) = \ 6X + \frac{1}{2} \cdot \left(1 - \frac{1}{2}\right) = 6X + \frac{1}{4}$$

$$U_1(C,C,D) = \pi_1(C,C) + \tilde{f}_2(C,D)\big(1 + \ f_1(C,C)\big) = \ 4X + \frac{1}{2} \cdot \left(1 + \frac{1}{2}\right) = 4X + \frac{3}{4}$$

$$U_1(D,C,D) = \pi_1(D,C) + \tilde{f}_2(C,D)\big(1 + \ f_1(D,C)\big) = \ 6X + \frac{1}{2} \cdot \left(1 - \frac{1}{2}\right) = 6X + \frac{1}{4}$$

$$U_1(C,D,C) = \pi_1(C,D) + \tilde{f}_2(D,C)\big(1 + \ f_1(C,D)\big) = \ 0 + \left(-\frac{1}{2}\right) \cdot \left(1 + \frac{1}{2}\right) = 0 - \frac{3}{4}$$

$$U_1(D,D,C) = \pi_1(D,D) + \tilde{f}_2(D,C)\big(1 + \ f_1(D,D)\big) = \ X + \left(-\frac{1}{2}\right) \cdot \left(1 - \frac{1}{2}\right) = X - \frac{1}{4}$$

$$U_1(C,D,D) = \pi_1(C,D) + \tilde{f}_2(D,D)\big(1 + \ f_1(C,D)\big) = \ 0 + \left(-\frac{1}{2}\right) \cdot \left(1 + \frac{1}{2}\right) = -\frac{3}{4}$$

$$U_1(D,D,D) = \pi_1(D,D) + \tilde{f}_2(D,D)\big(1 + \ f_1(D,D)\big) = \ X + \left(-\frac{1}{2}\right) \cdot \left(1 - \frac{1}{2}\right) = X - \frac{1}{4}$$

When $X < \frac{1}{4}$ we get the following with elementary algebra:

$$U_1(C,C,C) > U_1(D,C,C)$$

$$U_1(C,C,D) > U_1(D,C,D)$$

$$U_1(D,D,D) > U_1(C,D,D)$$

$$U_1(D,D,C) > U_1(C,D,C)$$

Therefore, if player 1 is a good type, then he will choose Act_1C if Bel_1Act_2C but he will choose Act_1D if Bel_1Act_2D. This is true independent of whether his second order belief is $Bel_1Bel_2Act_1C$ or $Bel_1Bel_2Act_1D$. The same reasoning applies to player 2 since the game is symmetric.

8 REFERENCES

Austin, John Langshaw (1975): How to do things with words: Oxford university press.

Axelrod, Robert (1980a): Effective choice in the prisoner’s dilemma. In *Journal of conflict resolution* 24 (1), pp. 3–25.

Axelrod, Robert (1980b): More effective choice in the prisoner’s dilemma. In *Journal of conflict resolution* 24 (3), pp. 379–403.

Axelrod, Robert; Hamilton, William D. (1981): The evolution of cooperation. In *science* 211 (4489), pp. 1390–1396.

Battigalli, Pierpaolo; Dufwenberg, Martin (2009): Dynamic psychological games. In *Journal of Economic Theory* 144 (1), pp. 1–35.

Benedict, Ruth (2005): The chrysanthemum and the sword: Patterns of Japanese culture: Houghton Mifflin Harcourt.

Bratman, Michael (1987): Intention, plans, and practical reason: Harvard University Press Cambridge, MA (10).

Cohen, Philip R.; Levesque, Hector J. (1990): Intention is choice with commitment. In *Artificial intelligence* 42 (2-3), pp. 213–261.

Connelly, Brian L.; Certo, S. Trevis; Ireland, R. Duane; Reutzel, Christopher R. (2011): Signaling theory: A review and assessment. In *Journal of management* 37 (1), pp. 39–67.

Darwin, Charles; Prodger, Phillip (1998): The expression of the emotions in man and animals: Oxford University Press, USA.

Dastani, Mehdi; Lorini, Emiliano (2012): A logic of emotions: from appraisal to coping. In : Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2. International Foundation for Autonomous Agents and Multiagent Systems, pp. 1133–1140.

Dastani, Mehdi; Meyer, John-Jules Ch; others: Programming agents with emotions. In : ECAI, pp. 215–219.

Falk, Armin; Fehr, Ernst; Fischbacher, Urs (2008): Testing theories of fairness—Intentions matter. In *Games and economic Behavior* 62 (1), pp. 287–303.

Farrell, Joseph; Rabin, Matthew (1996): Cheap talk. In *Journal of Economic perspectives* 10 (3), pp. 103–118.

Fehr, Ernst; Schmidt, Klaus M. (2006): The economics of fairness, reciprocity and altruism-experimental evidence and new theories. In *Handbook of the economics of giving, altruism and reciprocity* 1, pp. 615–691.

Fischer, Agneta H.; Manstead, Antony, SR; others (2008): Social functions of emotion. In *Handbook of emotions* 3, pp. 456–468.

Frank, Robert H. (1988): Passions within reason: The strategic role of the emotions: WW Norton & Co.

Frijda, Nico H. (1986): The emotions: Cambridge University Press.

Frijda, Nico H.; Mesquita, Batja (1994): The social roles and functions of emotions.

Geanakoplos, John; Pearce, David; Stacchetti, Ennio (1989): Psychological games and sequential rationality. In *Games and economic Behavior* 1 (1), pp. 60–79.

Gibson, James J. (1977): The theory of affordances. In *Hilldale, USA* 1, p. 2.

Griffiths, Paul E. (2004): Towards a Machiavellian theory of emotional appraisal. In *Emotion, evolution and rationality*, pp. 89–105.

Griffiths, Paul Edmund; Scarantino, Andrea (2005): Emotions in the wild: The situated perspective on emotion.

Haidt, Jonathan (2003): The moral emotions. In *Handbook of affective sciences* 11 (2003), pp. 852–870.

Horton, Thomas E.; Chakraborty, Arpan; Amant, Robert St (2012): Affordances for robots: a brief survey. In *AVANT. Pismo Awangardy Filozoficzno-Naukowej* 2, pp. 70–84.

Joffily, Mateus; Masclet, David; Noussair, Charles N.; Villeval, Marie Claire (2014): Emotions, sanctions, and cooperation. In *Southern Economic Journal* 80 (4), pp. 1002–1027.

Johnson, Matthew; Bradshaw, Jeffrey M.; Feltovich, Paul J.; Jonker, Catholijn M.; van Riemsdijk, Birna; Sierhuis, Maarten (2010): The fundamental principle of coactive design: Interdependence must shape autonomy. In : International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems. Springer, pp. 172–191.

Keltner, Dacher; Haidt, Jonathan (1999): Social functions of emotions at four levels of analysis. In *Cognition & Emotion* 13 (5), pp. 505–521.

Keltner, Dacher; Haidt, Jonathan; Shiota, Michelle N. (2006): Social functionalism and the evolution of emotions. In *Evolution and social psychology* 115, p. 142.

Klinnert, Mary D.; Campos, Joseph J.; Sorce, James F.; Emde, Robert N.; Svejda, MARILYN (1983): Emotions as behavior regulators: Social referencing in infancy. In : Emotions in early development: Elsevier, pp. 57–86.

Kraut, Robert E.; Johnston, Robert E. (1979): Social and emotional messages of smiling: an ethological approach. In *Journal of personality and social psychology* 37 (9), p. 1539.

Krueger, Joel (2012): Seeing mind in action. In *Phenomenology and the Cognitive Sciences* 11 (2), pp. 149–173.

Lazarus, Richard S. (1966): Psychological stress and the coping process.

Lazarus, Richard S. (1991): Emotion and adaptation: Oxford University Press on Demand.

Lelieveld, Gert-Jan; van Dijk, Eric; van Beest, Ilja; van Kleef, Gerben A. (2012): Why anger and disappointment affect other's bargaining behavior differently: The moderating role of power and the mediating role of reciprocal and complementary emotions. In *Personality and Social Psychology Bulletin* 38 (9), pp. 1209–1221.

Lerner, Jennifer S.; Li, Ye; Valdesolo, Piercarlo; Kassam, Karim S. (2015): Emotion and decision making. In *Annual review of psychology* 66.

Meyer, John-Jules Ch (2006): Reasoning about emotional agents. In *International journal of intelligent systems* 21 (6), pp. 601–619.

Neumann, John von; Morgenstern, Oskar (2007): Theory of games and economic behavior (commemorative edition): Princeton university press.

Ortony, Andrew; Clore, Gerald L.; Collins, Allan (1990): The cognitive structure of emotions: Cambridge University Press.

Osborne, Martin J.; Rubinstein, Ariel (1994): A course in game theory: MIT press.

Parkinson, Brian (1996): Emotions are social. In *British journal of psychology* 87 (4), pp. 663–683.

Pearlin, Leonard I.; Schooler, Carmi (1978): The structure of coping. In *Journal of health and social behavior*, pp. 2–21.

Pereira, David; Oliveira, Eugénio; Moreira, Nelma (2007): Formal modelling of emotions in BDI agents. In : International Workshop on Computational Logic in Multi-Agent Systems. Springer, pp. 62–81.

Picard, Rosalind W. (2000): Affective computing: MIT press.

Picard, Rosalind W. (2010): Affective computing: from laughter to IEEE. In *IEEE Transactions on Affective Computing* 1 (1), pp. 11–17.

Rabin, Matthew (1993): Incorporating fairness into game theory and economics. In *The American economic review*, pp. 1281–1302.

Rao, Anand S.; Georgeff, Michael P. (1991): Modeling rational agents within a BDI-architecture. In *KR* 91, pp. 473–484.

Raubal, Martin; Moratz, Reinhard (2008): A functional model for affordance-based agents. In : Towards affordance-based robot control: Springer, pp. 91–105.

Reisenzein, Rainer; Hudlicka, Eva; Dastani, Mehdi; Gratch, Jonathan; Hindriks, Koen; Lorini, Emiliano; Meyer, John-Jules Ch (2013): Computational modeling of emotion: Toward improving the inter-and intradisciplinary exchange. In *IEEE Transactions on Affective Computing* 4 (3), pp. 246–266.

Rendsvig, Rasmus; Symons, John (2019): Epistemic Logic. In Edward N. Zalta (Ed.): The Stanford Encyclopedia of Philosophy. Summer 2019: Metaphysics Research Lab, Stanford University.

Robbins, Philip; Aydede, Murat (2008): The Cambridge handbook of situated cognition: Cambridge University Press.

Rozin, Paul; Lowery, Laura; Imada, Sumio; Haidt, Jonathan (1999): The CAD triad hypothesis: a mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). In *Journal of personality and social psychology* 76 (4), p. 574.

Scarantino, Andrea (2017): How to do things with emotional expressions: The theory of affective pragmatics. In *Psychological Inquiry* 28 (2-3), pp. 165–185.

Searle, John Rogers (1969): Speech acts: An essay in the philosophy of language: Cambridge University Press (626).

Simon, Herbert A. (1967): Motivational and emotional controls of cognition. In *Psychological review* 74 (1), p. 29.

Simon, Herbert A. (1972): Theories of bounded rationality. In *Decision and organization* 1 (1), pp. 161–176.

Spence, Michael (1978): Job market signaling. In : Uncertainty in Economics: Elsevier, pp. 281–306.

Steunebrink, Bas R.; Dastani, Mehdi; Meyer, John-Jules Ch; others (2007): A logic of emotions for intelligent agents. In : Proceedings of the National Conference on Artificial Intelligence, vol. 22. Menlo Park, CA Cambridge, MA London AAAI Press MIT Press 1999, p. 142.

Tangney, June Price; Stuewig, Jeff; Mashek, Debra J. (2007): Moral emotions and moral behavior. In *Annu. Rev. Psychol.* 58, pp. 345–372.

van Kleef, Gerben A. (2009): How emotions regulate social life: The emotions as social information (EASI) model. In *Current directions in psychological science* 18 (3), pp. 184–188.

van Kleef, Gerben A. (2016): The interpersonal dynamics of emotion: Cambridge University Press.

van Kleef, Gerben A.; Dreu, Carsten K. W. de; Manstead, Antony, SR (2006): Supplication and appeasement in conflict and negotiation: The interpersonal effects of disappointment, worry, guilt, and regret. In *Journal of personality and social psychology* 91 (1), p. 124.

Wilutzky, Wendy (2015): Emotions as pragmatic and epistemic actions. In *Frontiers in psychology* 6, p. 1593.

Wong, Ying; Tsai, Jeanne (2007): Cultural models of shame and guilt. In *The self-conscious emotions: Theory and research*, pp. 209–223.