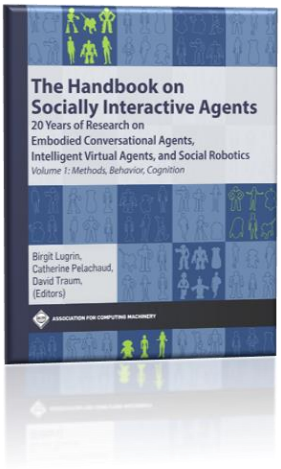
Empathy and Prosociality in Social Agents

Ana Paiva, Filipa Correia, Raquel Oliveira,
Fernando Santos, and Patrícia Arriaga



Author note:

This is a preprint. The final article is published in "The Handbook on Socially Interactive Agents" by ACM books.

Citation information:

Paiva, A., Correia, F., Oliveira, R., Santos, F., Arriaga, P. (2021). Empathy and Prosociality in Social Agents. In B. Lugrin, C. Pelachaud, D. Traum (Eds.), *Handbook on Socially Interactive Agents – 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics*, Volume 1: Methods, Behavior, Cognition (pp. 385-431). ACM.

DOI of the final chapter: 10.1145/3477322.3477334

DOI of volume 1 of the handbook: 10.1145/3477322

Correspondence concerning this chapter should be addressed Anna Paiva (ana.paiva@inesc-id.pt)

11 Empathy and Prosociality in Social Agents

Ana Paiva, Filipa Correia, Raquel Oliveira, Fernando Santos and Patricia Arriaga

11.1 Motivation

Although some scientists might disagree about the exact role and importance of emotions in our daily lives, virtually all people (including scientists) admit to occasionally experience sadness, joy and other emotions and can recognize how each one feels and how it affects them [Izard 2013]. From an evolutionary point of view, emotions carry significant advantages in terms of survival and, due to their universality, they allow for the non-verbal communication of inner psychological states and the easy recognition of those states in others. For this reason, many argue that emotions are essential features of complex social interactions and that they have adaptive functions [Ekman 1999, Gloria and Steinhardt 2016, Izard 2013]. So, it is not surprising the role that emotions play in the interactions between humans and technology, and in particular, social agents. Much of the work conducted in this area seems to clearly and repeatedly find that people interact with technological artefacts as more than mere tools; users often apply schemes for social and emotional interaction with other humans, to their interaction with machines (see Chapter 3 [Kramer and Manzeschke 2020]). For this reason, emotions are now a central part of the design, development and evaluation of new technologies [Picard et al. 2002], including social agents and robots. But, as we consider the development of emotional behaviours in social agents, as seen in the previous chapter (see Chapter 10 [Broekens 2020]), we also need to reflect upon the effects that emotions have on both humans and the agents as a result. We cannot detach how emotional behaviour experienced by one person affects another, and the responses that arise as a result. Indeed, many of our emotions are social and related to how the others feel, and empathy processes fit in this realm of responses. In general terms, empathy can be considered as the response to some other person's emotional state, where such response is more congruent to the others' emotional state than one's own. For example, if an agent reports a gloomy event, the human viewer may respond by feeling sad or even try to comfort or provide some advice to the agent.

Several efforts to create empathic agents have shown that the display of empathy can positively impact the user's feelings of trust and friendship towards such agents. Although an

interesting finding, one might ask why (and if) we should trust and nourish these emotional connections with technological items [Reeves and Nass 1996]. From an utilitarian point of view, the answer to this question is rooted in the assumption that positive feelings towards machines can lead to higher overall satisfactory and engagement in long-term interactions with technological devices or agents. This can have a positive impact in the commercial success of such devices, as modulated by users' intention to interact with them in the future and their attitudes and acceptance of such items in their daily lives. On the other hand, empathy, and empathic responses to agents, may also result in more awareness to someone else's emotions, fostering perspective taking and reasoning about others, often competencies we seek to promote.

We believe that our interactions with machines, and agents, can have a deep impact in our behaviors (both in actions directed at those machines and towards other humans). In particular, positive social behaviours (such as cooperation and prosocial behaviours) may be elicited through the interaction with socially interactive agents (SIA) that can invoke positive emotions from their users. These complex social behaviors, as many studies have proposed, can transcend the limited domain of interaction and lead to actual cooperation and prosocial behaviors directed at other humans.

In this chapter, we focus on empathy, prosociality, and the benefits that may accrue from such dispositions, both at the individual and societal level, when humans and agents interact. At an individual level, for example, prosocial spending seems to make individuals happier and result in higher levels of positive emotions [Dunn et al. 2014]. Similarly, individuals who engage in volunteering activities frequently report higher levels of happiness and health than those who do not [Borgonovi 2008]; and some authors have even observed a link between the offer of instrumental support to close family and friends (i.e. informal caretaking) with a lowered mortality rate [Brown et al. 2003]. In addition, engaging in prosocial behaviours seems to result in greater well-being to the prosocial actor due to its ability to satisfy the psychological needs of autonomy, competence and relatedness [Martela and Ryan 2016].

At a societal level, although hard to quantify, prosocial behaviours also bring many advantages. For example, in the United States of America alone, in 2018, it was estimated that around 30% of American citizens were engaged in some type of volunteering activity, which is the equivalent to over 75 million volunteers countrywide, whose total efforts account for a work and service effort valued at around 167 billion dollars. Volunteers are the foundation of essential organizations, such as the national disaster response system, which provides vital aid to the victims of hurricanes and other catastrophic events. However, the official number of volunteers does not include the many kinds of prosocial acts that are not considered formal volunteering, such as informal caretakers and people who perform prosocial actions towards friends and family. Moreover, the widespread movement of solidarity originated during the COVID-19 lockdowns is one example of how prosociality can act as a motivator for social support in hard times. Many different acts of kindness and help in the world were directed

at those in need, especially towards the elderly, care workers, and most often strangers. The use of hashtags such as #viralkindess were high and a sense of unanimity emerged as our unknown neighbours became our friends. The COVID-19 lockdowns were a time for empathy and prosociality. In different countries and continents, prosocial acts emerged such as giving free milk through a "kindness cooler" in Wisconsin, US, the help from cosmetic factories to produce and give away hand sanitizers, to the creation of community kitchens worldwide.

Moreover, [Pfattheicher et al. 2020] have shown that empathic concern for those most vulnerable to the COVID-19 predicted and promoted adherence to physical distancing and wearing face masks. These are two important behavioural measures recommended by the World Health Organization (WHO) to control the spreads of the SARS-CoV-2 contagion, which in turn contribute to facilitate health systems work and allow a better treatment to those infected. In the same line, [Campos-Mercade et al. 2020] have found the prosocial motivations was related to following these and other WHO health behaviour guidelines, and also to donating to fight COVID-19.

Given the benefits of prosocial behaviour, in this chapter, we will stand by the idea that, as we build agents that interact with humans, we need to go beyond social interaction and think about the effects that those agents can have in humans' well-being. We argue that effective socially interactive agents should not only be social, but also be prosocial: SIAs should be able to act in a prosocial manner and evoke prosocial behaviours from their users, directed at the agent, at other humans, and eventually at the society as a whole. The idea that interaction with SIA can contribute to the development of prosocial skills and interactions is not new. This idea is supported by psychological models of learning that propose that we learn and develop skills based on our interaction with other people and other social agents, in different contexts (e.g., direct interaction, playing games). In particular, according to the General Learning Model (GLM), people can extract information and learn from different situations and environmental interactions, through the employment of a wide range of cognitive mechanisms [Anderson and Bushman 2002, Barlett and Anderson 2012, Buckley and Anderson 2006]. This model attempts to explain how different life experiences can have an impact in a person's beliefs, attitudes, and cognitions [Gentile et al. 2009].

To address this challenge we build on previous work about empathy and prosociality in SIAs [Paiva et al. 2017] by providing a framework that accounts for the main variables that can be used to design prosocial agents, for both individual, group and society level interactions. This chapter makes a step in examining how empathy in the interaction between humans and agents can be achieved, and the role it plays in fostering prosocial and altruistic behaviour in general. This ultimate aim is the basis of the area of "prosocial computing", as initially described in [Paiva et al. 2018]. In this chapter, we first start by elaborating on the definition of the relevant concepts implicated in our work and by presenting a framework that captures both the potential effects of these concepts (empathy and prosociality), as well as the interactions among them, which are expected to produce prosocial behavior. Second,

we present an oriented and selective review of the literature regarding the currently existing models and architectures to build prosociality in agents followed by a review of user studies that have been conducted involving SIA. Finally, we present a selective review of prosociality models in populations and we conclude by outlining possible future avenues of research and discussion.

11.2 Concepts and Framework

Concepts such as prosociality, cooperation and altruism are important in many fields of psychology and other social sciences, as they underline the role that certain behaviours have in our daily lives resulting in important effects on how we behave towards each other and towards the society in general. For this reason, the search for the causes or antecedents that explain why people act prosocially, and what conditions facilitate that choice, has a long and fascinating history, that gathers multi-disciplinary contributions of many scientists from many areas of study yielding many interesting and sometimes even contradictory results. In particular, the pervasiveness of empathy, altruism, cooperation and prosociality in humans has for long puzzled biologists, economists, psychologists and researchers from multiple other disciplines [De Waal 2008, Hamilton 1964, Rand and Nowak 2013, Trivers 1971]. Although altruistic behaviours are commonly seen as some "heroic human acts", they are also observed in non-human species that exhibit complex social structures [Carter et al. 2017]. For example, in some species of birds, one can observe unrelated individuals protecting little fledgelings from predators, thus helping the breeding parents [Brown 1978]. Social insects (e.g., worker bees) give up their reproductive function in order to benefit their colonies [Hamilton 1972]. Some of these examples, where animals reveal apparent selfless behaviours, were a conundrum for Darwin: in a world where only the fittest survive, it is certainly puzzling that those sacrificing their own fitness – to benefit others – manage to win the contest of natural selection. Notwithstanding, Darwin himself advanced some explanations for the selection of altruistic behaviours, suggesting incipient notions of kin selection and reciprocal altruism. Some of these ideas were later elaborated. In the 60s, Hamilton developed ideas on kin selection, coining the today called *Hamilton's rule*. This rule postulates that altruistic cooperation evolves if the genetic relatedness between the cooperator and the recipient of the altruistic act, times the reproductive benefit gained by the recipient, outweighs the cost of altruism [Hamilton 1964]. Later on, Trivers formalized the idea of reciprocal altruism, proposing that altruistic cooperation can evolve if a cooperator helping today will be helped tomorrow [Trivers 1971]. Other mechanisms, such as indirect reciprocity, spatial selection and multi-level selection, were more recently studied [Nowak 2006, Rand and Nowak 2013]. These mechanisms can be seen as interaction structures that allow natural selection to choose, in the long-run, cooperative behaviours. In other words, these mechanisms constitute ultimate causes for cooperation. In parallel, research has advanced our knowledge on the proximate causes of cooperation, often rooted in psychological mechanisms. Empathy appears, in this context, as a prime

justification for altruism. In particular, the empathy-altruism hypothesis of Batson suggests that cooperation is triggered, regardless of the costs and benefits involved, if someone feels empathy towards another individual [Batson et al. 1995]. Similarly, Frans de Waal suggests that empathy is the ideal candidate mechanism that underlies altruism, especially altruism that arises in response to another person's pain, need and distress [De Waal 2008]. The mysteries of cooperation are not solved. In fact, explaining the evolution of cooperation was pointed out as a grand challenge for the XXI century [Pennisi 2005].

The vast scope of the factors involved in the study of prosociality requires on our part an initial clarification on different concepts required to describe the various approaches we will discuss in this chapter. They are:

- **Empathy** is defined by Hoffman [Hoffman 2001] as a psychological process that makes a person to have "feelings that are more congruent with another's situation than with his own situation". Empathy is a multidimensional concept usually distinguished in terms of cognitive and affective empathy [Davis 2018, Maibom 2017]. Cognitive empathy is the capacity to put oneself in the other position, by being able to see and understand what the recipient thinks and/or feels, also named perspective taking, and requires having a theory of another's mind (theory of mind). Affective empathy involves affect from the actor (the empathizer). Examples include "vicarious" affect, resonance or mirroring similar emotions of the recipient (also considered basic affective empathy). When applied to situations in which a recipient is in need or suffering, two different affective empathy dimensions have been proposed: empathic concern (also named sympathy or compassion) and personal distress. Empathic concern is the ability to feel other-oriented concerns, that is, sympathy for the welfare of others by resonating with others' negative emotions, and often gives rise to prosocial behaviours. Personal distress involves feeling distress for oneself (self-oriented concern) and for the recipient in need [Maibom 2017].
- **Prosocial behaviour** is a multidimensional concept that can broadly be defined as a voluntary behaviour intended to benefit another [Coyne et al. 2018, Eisenberg and Spinrad 2014]. Examples include altruism, solidarity, sharing, caregiving, comforting. It can vary from high cost (e.g. altruism, caregiving, volunteer, sacrificing) to very low cost behaviours (e.g. comforting), and is intimately related with other constructs such as cooperation, reciprocity, empathy, generosity, trust and fairness. The underlying motives to act prosocially can vary from being motivated to increase another's welfare (other-oriented) to increase one's welfare (self-oriented) [Eisenberg et al. 2016]. When there is no expectation of self-gain the behaviour is considered altruistic, but when enacted because of the request of others or internalised social norms, it is associated with compliance [Xiao et al. 2019], suggesting moral reasons such as gratitude [Ma et al.

2017]. Specific emotions may also play an important role in prosocial actions. Emotions often associated with prosocial behaviours include sympathy, compassion, guilt, regret, but their role is also highly dependent on the type of prosocial behaviour, underlying motivation(s), context, individual differences, and group factors.

- **Altruism** is an unconditional prosocial tendency for an agent to act to benefit the recipient and increase his/her welfare [Batson 2011] without the expectation of any self-gain (thus opposed to egoism) [Van Lange et al. 2014]. However, while some authors consider that altruism does not preclude the agent from benefiting from the behaviour, other authors argue that even altruistic actions benefit the agent in some way. For this reason, they sustain that there is no such thing as authentic, genuine, or "true" altruism. In contrast, for other authors, authentic altruism occurs when the altruistic action has some cost to the agent's personal interest (for a review [Schramme 2017]). In fact, in biology, altruism often refers to behaviours which are costly (in terms of reproductive fitness) to an organism and beneficial to the recipient [West et al. 2007]. Similarly, in economics, altruism is defined as costly behaviours that confer economic benefits on other individuals [Fehr and Fischbacher 2003].
- **Reciprocity** is defined in broad terms as "treating another in the same way as the one is treated" [Kolm 2008]. Many forms of reciprocity have been described, but the most common is direct and indirect reciprocity. Direct reciprocity involves a mutual and direct exchange between an agent (A) and a recipient (B). Indirect reciprocity occurs when the reciprocal acts involve another person (C) who is not the initial recipient (B) and can be divided into upstream and downstream reciprocity. Upstream reciprocity occurs when the agent (A) acts prosocially towards a person (B) after receiving some prosocial behaviour from the recipient (C). Downstream reciprocity corresponds to an increase in the likelihood that the agent (A) will be a recipient of prosocial behaviour from another person (C) after acting prosocially towards a former recipient (B), and this likelihood is expected because it benefits the agent reputation [Ma et al. 2017, Nowak and Roch 2007]. Reciprocity – both direct and indirect – have been pointed as fundamental mechanisms to explain the evolutionary origins of altruistic cooperation [Nowak 2006, Nowak and Sigmund 1998, Rand and Nowak 2013, Trivers 1971].
- **Cooperation** is a type of prosocial behaviour involving efforts to enhance joint positive outcomes for both the agent and recipient(s). However, cooperation has also different forms depending on the motivation. The most common distinction is between instrumental and non-instrumental (or elementary) cooperation. Instrumental cooperation refers to all behaviour by which individuals contribute to the quality of a system that rewards cooperation and punishes non-cooperators. Yet, that actions are performed as a mean to obtain self-benefit, that is, the agent performs the cooperative action(s) because it will enable achieving certain outcomes, including positive outcomes for the recipient(s).

Cooperation is also often referred to, in biology, as behaviour that provides benefits to another individual [West et al. 2007]. In this regard, one may distinguish between altruistic cooperation – referring to altruism, defined above – and collaboration [Tomasello and Vaish 2013] or mutualism [West et al. 2007] – when both the agent and the recipient benefit from the cooperative relation. Across different disciplines [Fehr and Fischbacher 2003, Rand and Nowak 2013, Wu et al. 2020], *cooperation* has been used to refer to altruistic cooperation, that is, costly behaviour that confer benefits to other individuals.

- **Selfishness** is considered the motivation for self-benefit (egoistic or self-oriented concern) without concern for others interests and well-being [Crocker et al. 2017]. It underlies most of the current approaches to the creation of "rational agents" [Wooldridge 2003] very much inspired in the *homo economicus* notion. The idea that humans act in pure self-interested way, trying to optimize their gains disregarding the other's welfare, as adopted in many economic theories, is the root for many approaches for designing rational agents [Wooldridge 2003]. However, humans do not act in a completely selfish way, as many studies involving social dilemmas played by humans have shown. Instead, humans cooperate and act altruistically at their own personal cost. For example, in the well-known prisoner's' dilemma, where the rational strategy is to defect, it was found through a meta-analysis that humans on average cooperated 47.4 % (cooperation rate) [Sally 1995].

11.2.1 From empathy to prosociality

According to Hoffman [Hoffman 2001] empathy is the "spark of human concern for others, the glue that makes social life possible", underlying the strong effect that empathy has towards the establishment of social bonds. More specifically, empathy has been widely thought of as an "other-centered" emotion, that facilitates the understanding of other people's situations, by allowing us *to put ourselves in another person's shoes* (i.e. perspective-taking) [Batson et al. 1991, Rumble et al. 2010]. So it is not surprising that many studies have reported positive relations between empathy and prosocial behaviour. In particular, Batson, by means of a set of experimental designs, tested the hypothesis that empathy (in particular, empathic concern) is a strong predictor of altruistic motivation and behaviour, also known as the empathy-altruism hypothesis [Batson 2011, 2014]. In human-human interactions, empathy has been linked to a number of prosocial behaviours such as helping and cooperating in contexts in which such behaviours do not serve the individuals' immediate selfish objectives [Batson and Ahmad 2001, Batson et al. 1995]. Empathy is thought to sustain these type of prosocial behaviour by increasing the positive weight assigned to the other's outcomes, consequently increasing generous behaviour from the first to the latter [Rumble et al. 2010].

Regarding empathic-related traits, a recent meta-analysis on the predictive role of personality on prosocial behaviour across several interdependent situations [Thielmann et al. 2020]

has shown that one of the strongest positive predictors of prosocial behaviour were the traits of unconditional concern for others.

In fact, many studies priming empathic concern towards a recipient (usually framed as a victim) showed more altruism, even when these behaviours were against the person's personal interest [Batson 2014]. Less consistent is the role of personal distress on prosocial behaviour. Although personal distress also tends to co-occur with empathic concern towards a recipient expressing distress, actions vary depending on how much distress the recipient feels. When the distress is experienced as overwhelming, it is often associated with a tendency to withdrawal from the distressing context, thereby compromising prosocial acts towards the recipient in need [Maibom 2017]. In spite of this, affective empathy, and empathic concern in particular, seems to be one of the main predictors of prosocial behaviour.

However, cognitive empathy also seems to play an important role. For example, in three studies, Galinsky at al. [Galinsky et al. 2008] have shown that perspective taking (understanding others' interests and motives) was more useful in negotiation processes than affective empathy. Thus, it is important to understand which empathic dimension and under which circumstances they arise, to establish the relationship with prosocial behaviours. Furthermore, when agents are mixed in these types of interactions, all these dimensions need to be articulated and somehow engineered.

11.2.2 Prosocial Agents: dimensions of the current analysis

The goal of this chapter is to provide an overview of the area of socially interactive agents (SIAs) that act in situations where empathic processes and prosocial behaviours exist. Thus, we should consider the multitude of roles and situations where agents can participate in and the different ways by which humans may interact with them. As a way to characterize these scenarios, let us consider that there are two humans/agents: a *recipient* and a *subject*. The *recipient* is the agent experiencing an emotionally charged situation (for example when one is given some bad news) and potentially expressing it to others. This expression can be displayed through facial expressions or even by uttering the sentiment felt in natural language. As a result, the *subject* responds to the situation and the feelings of the *recipient* experiencing an empathic response (see empathic phase in Figure 11.1), and eventually acting in a prosocial manner (see the prosocial phase in Figure 11.1).

We can say that we are witnessing prosocial behaviour when the *subject* incurs some cost (Cs) as he/she acts to provide some gain to the *recipient* (Gr). As mentioned before, the *subject*, him/herself may also obtain some gain from the prosocial action. In fact, in many scenarios, that is the case, as the positive effects of prosociality are enormous as already mentioned.

The characterization of the roles that SIAs and humans play in this analysis framework allows for the following possibilities:

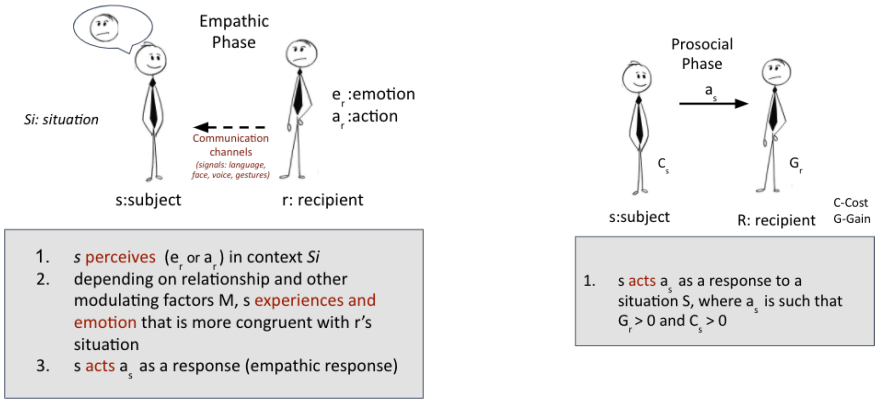


Figure 11.1 Generic situations of empathic and prosocial behaviours between two agents

- SIAs that act as the *subject* in the empathic phase: the agent is in a situation that perceives others (agents or humans) and its internal mechanisms allows for it to respond empathically;
- SIAs that act as a *subject* in the prosocial phase: the agent acts towards the *recipient* in a prosocial manner;
- SIAs that act as *recipients* in an empathic phase: agents act in scenarios to evoke empathy in others (including users)
- SIAs that act as *recipients* in an prosocial phase: the agents promote prosocial behaviours in others

These different types of roles for the agents require from them a myriad of design features and computational processes. An agent acting as the *subject* needs to be equipped with mechanisms that allow it to perceive and appraise the situation, be able to reason about the others, and respond adequately. Features such as emotional recognition or perspective taking may be essential for SIAs to act as *subjects*, but not necessary for acting as *recipients*. Other features such as the SIA's embodiment (disembodied, virtually embodied or physically embodied) may also be more important in one type of context than another. For example, a SIA acting as a *recipient*, may use aspects of its embodiment (its gaze, lights, posture) that may be vital to convey the emotional state in a situation.

The roles of SIAs can be extended when we move from traditional dyadic interactions, as portrayed in Figure 11.1 to groups featuring both humans and agents (see Chapter 17 of this book and [Correia et al. 2018b]). In many situations we can also have agents acting as bystanders witnessing empathic and prosocial situations. For example, agents may operate

in a group context where other agents or humans act in a manner that fires some emotional and empathic responses. Creating agents as bystanders can be inspired in the BIM model (Bystander Intervention Model) proposed by Latané and Darley [Latané and Darley 1970] that was developed to examine bystander behaviour in emergency situations. This model describes a set of successive phases which an individual must experience to intervene in a situation, namely, the perception of the event, the interpretation of the degree of emergency of the event, the recognition that it is the agent's responsibility to intervene, to knowing what to do and intervening. This model has actually been observed to characterize the behaviour of bystander adolescents in cyberbullying cases [Ferreira et al. 2020a], opening doors to agent based interventions with agents acting as bystanders.

This move from dyadic interactions to groups and societies is of paramount importance when analysing the role that SIAs may have in real-world scenarios. As proposed by Penner et al. [Penner et al. 2005] the analysis of prosocial behaviour can be done at three different levels: the micro-level, the meso-level and the macro-level. At a micro-level, the study of prosociality is done around the origins of prosocial tendencies in general and sources of variation for these tendencies. In the meso-level, the study is done around helper-recipient situations (similar to what is shown in Figure 11.1). Finally, at the macro-level, prosociality is studied in the context of groups and societies. These three levels are interconnected, and if we place agents to interact with humans, we need to consider different levels of analysis. In this chapter, we will explore the role of empathy and prosociality in social agents along three different levels (see Figure 11.2). The first one is the individual level **(A)**, where we will detail the internal processes that lead to empathy and prosociality, and how those processes can be integrated and engineered in SIAs (see section 11.3). The second level **(B)** is the interaction level where we examine at dyadic interactions and where we review the mechanisms and processes that affect how humans interacting with socially interactive agents, particularly focusing on the effects that SIAs have on human prosociality (section 11.4). Finally, we believe that prosociality in human-agent interactions needs to be examined beyond isolated encounters, that is, embedded in dynamic populations and acknowledging possible long-term effects. Therefore, at the third level of analysis **(C)** we will explore the role of prosociality at the macro-level, that is, in groups and in (hybrid) populations of humans and social agents (section 11.5).

11.3 Models and Architectures to Build Empathy and Prosociality

In order for SIAs to act in empathic and prosocial situations they need to be equipped with computational mechanisms that include perception, decision making an action execution, underpinning a traditional agent modelling approach. Moreover, models and architectures to create empathic agents are also inspired by existing theories of empathy in humans providing ways to identify and structure the computational processes in social agents. In general terms these theories may follow two distinct approaches: on the one hand, *categorical* approaches carefully distinguish the affective empathy from cognitive empathy; on the other

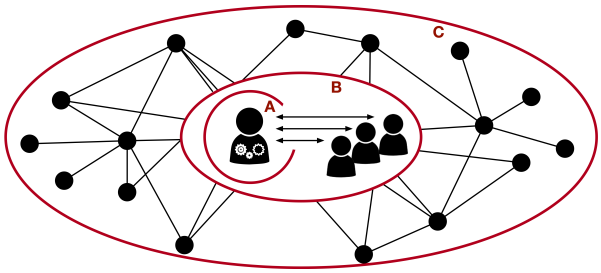


Figure 11.2 Dimensions of analysis: A - agent architectures: focus on the internal processes that lead to empathy and prosociality, and how those processes can be integrated and engineered in social agents; B - social agents: focus on the mechanisms and processes that affect how humans interact with social agents, particularly focusing on the effects that SIA have on human prosociality; C - social agents within populations: focus on the role of SIA in stabilizing prosociality in (hybrid) populations of humans and agents.

hand, *dimensional* approaches propose that both affective and cognitive mechanisms can be integrated into a multidimensional system. This implies that architectures may feature different computational processes accordingly. In a recent survey, Yalçın and DiPaola have systematically analysed the literature on empathic agent architectures by separating affective mechanisms, also refereed as low-level functions, from cognitive mechanisms, also referred as high-level functions [Yalçın and DiPaola 2019b]. Their framework and, in particular, the distinction between affective and cognitive mechanisms of empathy aims at highlighting the similarities and overlaps between the existing models and how some functions of these models can be functionally integrated.

In contrast, the framework based proposed by Boukricha et al. [Boukricha et al. 2013], and extended in [Paiva et al. 2017], proposes the following components in a general architecture: (1) **empathy mechanisms** —"the process by which an empathic emotion arises"—; (2) **empathy modulation** —"the process by which both an empathic emotion is modulated and a degree of empathy is determined"—; and (3) **empathic responses** —"the process by which an empathic emotion is expressed/communicated and actions are taken" [Paiva et al. 2017]. This framework specifically aggregates low- and high-level functions into the empathy mechanisms, as well as it merges their outputs into empathic responses. In other words, it acknowledges that empathic responses by artificial agents may still occur regardless of the mechanisms behind having an affective and/or a cognitive process(es). We are particularly interested in this framework as we postulate that an empathic response by artificial agents, independently from their theoretical and methodological approach, can lead to prosocial behaviours (see Fig. 11.3).

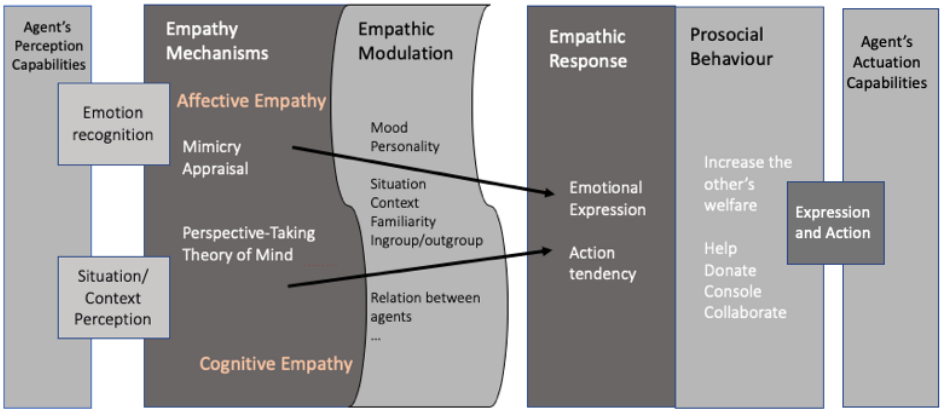


Figure 11.3 Empathy Processes and Mechanisms in SIAs leading to Prosocial Behaviour

Finally, to bridge artificial empathy and prosociality in SIAs, we will first overview existing architectures and computational models to create empathic agents and then discuss how artificial empathy may lead to prosocial behaviours. We will base this connection in Batson's hypothesis that empathy (in particular, empathic concern) is a strong predictor of altruistic motivation and behaviours. However, from an architectural point of view, we assume that agents must have different cognitive and affective mechanisms, often inspired by humans. Thus, perception, cognition, motivation, emotions, and interpersonal behaviours ought to be considered as they are essential for creating intelligent and social behaviour in agents.

11.3.1 Empathy mechanisms for Empathic Agents

Empathy mechanisms constitute the internal processes that lead to an empathic emotion to arise. Hence, empathy mechanisms for artificial agents are closely related to their perceptive skills. Generally, empathic agents are required to be aware of others, either by recognizing their emotions or their actions within a certain context, from which they can then infer or interpret the others' goals, intentions and/or affective state. Additionally, empathy mechanisms may also depend on the modalities that agents use to interact, which in turn may increase the complexity to model empathy. For instance, there are rule-based systems in which the agents are able to produce empathic behaviours, such as sympathetic or encouraging utterances, by interpreting the context or by performing a situational appraisal [Becker et al. 2005, Bickmore and Picard 2005, Leite et al. 2014, Lisetti et al. 2013, Prendinger and Ishizuka 2005]. Other more complex behaviours may present sophisticated models or architectures according to different methodological approaches.

One of the most used approaches is the analytical or theory-driven, in which computational models are based on theoretical models of empathy established in psychological and neuropsychological research. Mimicry is considered a fundamental mechanism for empathy and is supported by both the perception-action hypothesis [Preston and De Waal 2002] and the shared affective neural networks [De Vignemont and Singer 2006]. Some examples have explored such affective mechanisms using motor mimicry [Gonsior et al. 2011, Riek et al. 2010], or affective matching techniques [Boukricha et al. 2013, Leite et al. 2014, Lisetti et al. 2013]. Regarding cognitive mechanisms, four works should be highlighted. Firstly, the use of perspective-taking through self-projection both in [Leite et al. 2013] and [Rodrigues et al. 2015]. On the one hand, Leite et al. have used the appraisal mechanism of the robotic agent to appraise its companion's situation for that particular game context [Leite et al. 2013]. On the other hand, Rodrigues et al. go a step further by proposing a general model for the agent to appraise the target's situation using its own belief system and goals to appraise the other's situation as its own [Rodrigues et al. 2015]. Similarly, Boukricha et al. used regression functions that map the activation of Action Units into the Pleasure-Arousal-Dominance space as a shared representational system, (i.e. to both to animate the agent and to infer the emotional state of other agents). Recently, Yalçın & DiPaola proposed another computational model of empathy [Yalçın and DiPaola 2018], which is inspired by the Russian Doll model of empathy [De Waal 2007]. This last approach not only allows for other-oriented perspective-taking, such as theory of mind, but also an isolated information processing between low- and high-level mechanisms of empathy.

Another methodological approach to create computational models of empathy is the empirical or data-driven, in which the models are obtained from collected data and constitute generalisations of empathic behaviours and/or empathic situations. Within this methodological approach, McQuiggan et al. used data from human-human social interactions in a virtual environment to create two classifiers: one to learn *when* and another to learn *how* the agent can act empathetically [McQuiggan et al. 2008]. Similarly, Ochs et al. created a model to express empathic emotions based on an empirical analysis of human-agent dialogues [Ochs et al. 2012]. The collected dialogues were annotated according to their conditions of elicitation, which matched the theoretical appraisal theory they have used [Scherer 1988].

The last methodological approach is considered hybrid as it includes both empirical and theoretical processes for the agent to learn and/or express empathy. Within this methodological approach, we would like to highlight two works which both follow a developmental perspective. Firstly, Lim et al. explored the learning process of mirroring mechanisms as an emergent empathic behaviour and, therefore, their work was mostly focused on low-level empathy [Lim and Okuno 2015]. On the other hand, the work by Asada et al. proposed that empathic development can emerge from a parallel between imitation (such as motor mimicry or emotional contagion) and other cognitive mechanisms (such as self-other distinction) [Asada 2015]. Another example can be found in the Emote project [Alves-Oliveira et al. 2019] where

an autonomous robot was designed with empathic competencies to foster collaborative learning in adolescents in particular towards sustaining positive educational outcomes in long-term collaborative learning. The computational model built to drive the behavior of the SIA (embodied as the NAO robot) was a "hybrid behavior controller" combining a rule-based component and a data-driven one. The data-driven component was built with a dataset created using a restricted perception Wizard-of-Oz study [Sequeira et al. 2016]. The final system was tested in the robot showing its capability to foster meaningful discussions among students interacting with the robot and among themselves.

11.3.2 Empathy modulation in Empathic Agents

Empathy modulation is the process by which the empathic emotion or the degree of empathy are shaped by features of the agents and the situation. Empathy modulation is inherently coupled to the empathy mechanisms, as it shapes them, changing the result of the process. This modulation reflects in humans the individual differences found, as well as the type of relationship that exists between subject and recipient. Paul Bloom discusses negative effects of empathy due to modulation, arguing that "empathy is biased", and may "push us in the direction of parochialism and racism" [Bloom 2017]. However, despite of the importance of this aspect for studying empathy, so far only a few computational models of empathy have included modulation factors in their architectures. For instance, Mcquiggan et al. have considered in their data-driven model the following features of the observer: gender, age, user empathetic nature, and goal orientation [McQuiggan et al. 2008]. Boukricha et al. have not only included features of the *subject* (observer), such as the mood, but also liking and familiarity to represent the social relationship with the *recipient*, as well as the desirability of the observed emotion [Boukricha et al. 2013]. Similarly, the model proposed by Rodrigues et al. supports the following modulation factors: mood, personality, affective link and similarity [Rodrigues et al. 2015] (see Figure 11.4). Another factor that modulates the empathic responses is the strength of the emotional situation, the context, and the valence and intensity of the emotions exhibited by the target. These different categories of empathy modulators for computational models of empathy [Paiva et al. 2017] need further investigation, in particular in what concerns situational or context-related factors.

11.3.3 Empathic responses in Empathic Agents

Empathic responses can include both the expression of attitudes as well as actions and action tendencies. In fact, prosocial acts can result from an empathic emotion. Furthermore, the expression of empathy in SIAs can be displayed through different channels or modalities, according to the social affordances of the agent. The most common empathic response in SIAs is body expression [Riek et al. 2010] and, in particular, facial expression [Becker et al. 2005, Bickmore and Picard 2005, Boukricha et al. 2013, Lisetti et al. 2013, McQuiggan et al. 2008, Ochs et al. 2012, Rodrigues et al. 2015, Yalçın and DiPaola 2019a]. This means

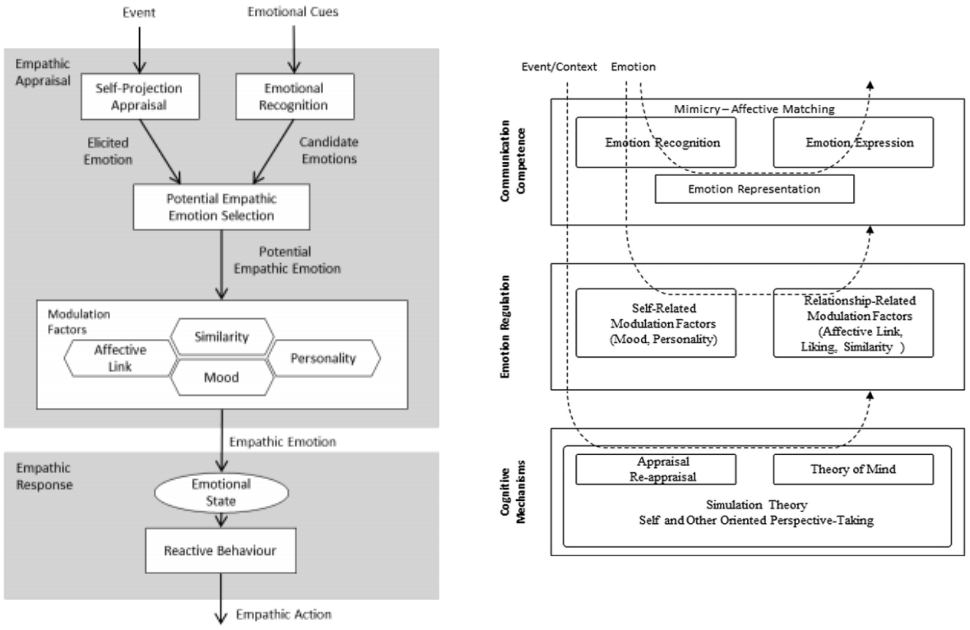


Figure 11.4 Examples of Architectures for Empathic SIAs taking modulation into account: (a) from [Rodrigues et al. 2015] and (b) from [Yalçın and DiPaola 2018].

that the body of the SIA must include some form of "face" or eyes. A few examples have also conveyed the empathic emotion on conversational settings through language [Bickmore and Picard 2005, Brave et al. 2005, McQuiggan et al. 2008, Prendinger and Ishizuka 2005]. Finally, an additional empathic response identified by Paiva et al. is the action tendency [Paiva et al. 2017], which is the readiness or urge to carry a behaviour upon a certain stimulus being prompted. In the next section, we will discuss existing theories that address how action tendencies and empathic responses, in general, may precede prosocial behaviour.

11.3.4 From Artificial Empathic Responses to Prosocial Behaviour

According to Batson et al. the major source of altruistic motivation is empathy, an other-oriented emotional response. This emotion is "elicited by and congruent with the perceived welfare of a person in need" [Batson et al. 2015] and is frequently reported as pity, compassion, or sympathy. Note that not all empathic emotion leads to altruistic motivation. For example, one may feel joy for another that received some good news, and that is still considered an empathic response. However, altruism and prosociality result from empathy felt when another is perceived to be in need. The empathy–altruism hypothesis claims that empathic

concern is one of the main drivers of altruistic motivation, and thus conducent to prosocial actions.

Recently, Costantini et al. proposed a simulation model of prosocial behaviours that integrates both descriptive and normative approaches [Costantini et al. 2019]. When analysing literature that explores basic processes and determinant variables of prosocial behaviours, the authors distinguish between descriptive-emotive approaches and normative-evolutionary approaches. While the first ones mainly aim at explaining the psychological motivations of prosociality, the latter ones look for an evolutionary benefit on prosocial acts. We will leave the discussion of the normative-evolutionary approaches to the meta-level analysis of social agents in societies (see Sec. 11.5). Thus, in the same vein as Batson [Batson et al. 2015] and Costantini et al. [Costantini et al. 2019] that role of emotions and empathy is prominent as an antecedent of prosocial behaviours. Nevertheless, so far, little work has been done in reflecting this link into the SIAs community, and in particular in agent's architectures. Considering the models and architectures we have previously reviewed, they may present distinct mechanisms to produce an empathic response, but some of them may even trigger hierarchically more than one empathic response [Yalçın and DiPaola 2018]. In both cases, regardless of the particular mechanism(s) being used, *how can an agent act prosocially upon having an empathic response?*

One concrete framework from psychology, the SAVE framework (Sociocultural Appraisals, Values, and Emotions), is a good example to draw the first steps for creating prosocial behaviour, as it provides an equation that tries to mirror the complex deliberative processes that occur during prosocial decisions by humans [Keltner et al. 2014]. If such an equation can be used by an agent to consider prosocial acts, empathy mechanisms can also be used to calculate some of the parameters. Cognitive mechanisms can contribute to infer the benefits of an action for someone else, referred as $B_{recipient}$. Similarly, empathic agents might also determine their own benefit of performing prosocial behaviours, referred as B_{self}, upon their empathic responses. For instance, the negative relief theory [Baumann et al. 1981] suggests that helping behaviours can reduce negative states of its actor.

11.4 Empathy and Prosociality in the Interaction with SIAs

Within the multitude of environmental factors that can have an impact on the interaction with technology, empathy and prosociality seem to be gaining increasingly more relevance. At the same time, the use of SIAs is now more widespread, as online virtual interactions turn ever more common, virtual characters begin to be used more and more in different applications, and social robots start to gain terrain as actors in our daily activities. As such the question of *if* and *how* these SIAs interact and affect people's behaviours, both in real-life and virtual scenarios, has received a considerable amount of attention recently.

While in the previous section we delved into the internal computational mechanisms that are required for empathic and prosocial SIAs, in this section, we consider how empathic and

prosocial interactions between humans and SIAs unfold (see Figure 11.2). So, we will discuss some of the issues underpinning the different factors that affect the emergence of empathic and prosocial responses in the interaction between humans and SIAs.

11.4.1 Research and application scenarios

The concept of prosociality is important in many fields of psychology, biology and economics, as it underlines many behaviours that are central components of our daily lives and have important effects on how we behave towards each other and towards the society in general. For this reason, the search for the causes or antecedents that explain why people choose to act prosocially, and what conditions facilitate that choice, has a long and fascinating history, that gathers the multi-disciplinary contributions of many scientists and yields many interesting results. As such, many studies have attempted to determine the factors that influence human prosociality not only in human-human scenarios but also towards the agents and towards other humans in virtual spaces. Studies in this area make use of an array of social games or dilemmas to model important aspects of social interaction and prosocial behaviours, such as the *Prisoner's Dilemma*, *Trust Game* [Berg et al. 1995] or *Public Goods Games* (see [Gotts et al. 2003] [Dawes 1980] and [Van Lange et al. 2013]). Social dilemmas can be broadly defined as situations in which short-term self-interest is at odds with longer-term collective interests [Van Lange et al. 2013], and they are particularly important as research settings to explore the strategic interactions between agents. In the context of social and prosocial interaction studies, they have been widely used given that they usually present a scenario in which participants are asked to decide between taking a selfish course of action, that serves their own immediate interest, or a prosocial course of action, that serves the collective interests. The flexibility and widespread use of these games have allowed for the emergence of a vast, interdisciplinary body of research (see [Berg et al. 1995] [Sally 1995] [Tavoni et al. 2011]) [Rand et al. 2012] [Rand and Nowak 2013]), that has, to some extent, provided counter-evidence to the thesis of *Homo Economicus* (i.e. the argument that people are mostly guided by external, selfish individualistic interests) [Gotts et al. 2003]. In addition, these research settings grow in relevance as they can be used to represent and model several collective, complex group situations, that are often dependent of the actions of large groups of independent agents (e.g., climate change or resource depletion) [Gotts et al. 2003] and humans, as will be discussed in Section 11.5. Furthermore, this type of scenarios constitutes a basis for the study of what are the characteristics (e.g., behaviour, embodiment) of social agents that can be manipulated to foster prosocial behaviour.

Despite the widespread use of social dilemmas as settings for studying empathy and prosociality, they are nevertheless artefacts where the complexity of real-world cases is reduced, allowing for researchers to pinpoint the exact elements to study and draw conclusions from very controlled situations. However, they are often too simple for real-world applications. Real-world scenarios are messier and involve many more variables, but SIAs may offer the

potential for change in real-life interactions with humans. As examples consider the use of a SIA acting as an empathic virtual nurse to promote behaviour change in health-related issues [Bickmore and Picard 2005], or the recent work by Morris et. al. [Morris et al. 2018] that created an empathic conversational agent to help people with mental health problems. In fact, health and education are application domains where the use of empathic SIAs has been quite prominent. In the particular case of education applications, SIAs can be used in several topics and for distinct target users. For example, the TARDIS system is an example where a SIA is used to coach young adults in the context of job interviews [Anderson et al. 2013]. SIAs have also been used in a game CRYSTAL ISLAND in the domain of microbiology for middle school students interviews [Sabourin et al. 2011] or embodied as robots exhibiting aspects empathy processes to train children and adolescents to understand geography and sustainability [Castellano et al. 2013] [Alves-Oliveira et al. 2019]. In fact, agents can also be used to foster the development of prosocial skills, and many interventions aimed at triggering prosociality have been developed in the past few years, with varying degrees of success [Goldstein et al. 1994, Leiberg et al. 2011, Lukinova and Myagkov 2016, Schellenberg et al. 2015]. Some of these interventions have become technology-based [Ibrahim and Ang 2018] opening doors for real world cases for SIAs. Some of these interventions are discrete in time and highly targeted at providing intentional prosocial training (e.g., [Lukinova and Myagkov 2016], where others seek to invoke prosocial skills in a more continuous manner through the interaction with computerized agents. In addition, as demonstrated by the study conducted by Kozlov and Johansen [Kozlov and Johansen 2010], prosocial behaviours in virtual environments seem to obey to the same influences as prosocial behaviour in real-life environments. That is, virtual environments constitute a good setting to explore some of the human-to-human studies on empathy and prosociality. For example, the existence of a large group of bystanders and the imposition of time constraints to help both seem to hinder participants' helping behaviours towards virtual agents [Kozlov and Johansen 2010]. This transference of the psychological determinants of prosocial behaviour and the subsequent prosocial responses within virtual environments (and towards social agents) falls in line with the predictions of the media equation theory [Reeves and Nass 1996], which broadly states that technology that can elicit social responses from humans, similar to those elicited by other humans the same social situations. These can include, much like in-person prosocial behaviour, user-related variables (such as personality [Graziano et al. 2007, Habashi et al. 2016, Hilbig et al. 2014, Pursell et al. 2008], dispositional compassion and empathy [Lim and DeSteno 2016, Lupoli et al. 2017, Rameson et al. 2012] or emotions [Batson 2014]), virtual environment-related variables (such as the presence of bystanders [King et al. 2008, Kozlov and Johansen 2010, Slater et al. 2013]) and agent-related variables (such as ethnicity [Gamberini et al. 2015] and gaze behaviour [Slater et al. 2013]).

In other words, exposure to prosocial content in virtual environments (often with the presence of SIAs) is expected to have both short-term (e.g., by increasing positive affect

[Saleem et al. 2012]) and long-term impacts (e.g., through changes in trait empathy [Prot et al. 2014]) in people's prosocial behaviours, motivations and tendencies [Coyne et al. 2018]. For example, Gentile demonstrated that repeated engagement with prosocial games can result in players transferring and generalizing prosocial motivations in real-life scenarios that are similar to those presented in the game, consequently resulting in greater helpful and cooperative behaviours [Gentile et al. 2009]. This effect seems to be culturally robust and to remain stable across different age ranges and levels of exposure (long-term and short-term) [Gentile et al. 2009, Greitemeyer and Mügge 2014, Saleem et al. 2012]. Recently Ferreira et a. [Ferreira et al. 2020b] examined whether experiencing a multiplayer serious game could foster cognitive empathy and prosociality in adolescent bystanders of cyberbullying (see Figure 11.5). The game uses SIAs acting as victims, bullies and bystanders in the game, and the results suggest an effect in increasing prosociality when compared with a control group. In a similar context the FearNot! game [Aylett et al. 2005] was developed to foster empathy towards a victim of bullying and promote behaviour change (see Figure 11.5). The game, featuring SIAs in a storytelling environment was designed to help children experience effective strategies for dealing with bullying. The results of a large scale evaluation showed a short-term effect on escaping victimization for a priori identified victims [Sapouna et al. 2010].

Indeed, studies using virtual reality as a method to create more lifelike and ecological valid scenarios to observe prosocial behaviour, suggest that prosociality can be elicited through a number of factors. For example, one study that manipulated the affordances (super-hero flight or riding as a passenger in a helicopter) given to players in the context of a simulated search and rescue activity, showed that participants given a super-power were more likely to engage in real-life prosocial behaviour immediately after the study [Rosenberg et al. 2013]. These results are in line with previous studies priming super-hero concepts to influence prosocial behaviour, which have found that priming was effective not only at increasing participants' immediate likelihood of helping in hypothetical situations, but also their engagement in prosocial activities (specifically, volunteering) three months after [Nelson and Norton 2005].

Given the distinct research scenarios and different areas that take advantage of SIAs to support learning or promote of prosocial behaviours and attitudes, the question that arises is what specific factors can contribute to elicit such behaviours a still needs further development.

11.4.2 Agent's characteristics, empathy, prosocial outcomes and measures

Although research on prosociality in the context of SIA is still quite new, many studies have already been conducted to investigate which agents' characteristics can impact empathic responses from users and nudge them towards prosocial courses of action. In fact, some studies have shown that empathy in SIAs seems to have a positive impact in cooperation and prosocial behaviour, a result that is in line with the empathy-altruism hypothesis [Baumann et al. 1981]. Nevertheless, in agents the display of empathy requires the agent to be able

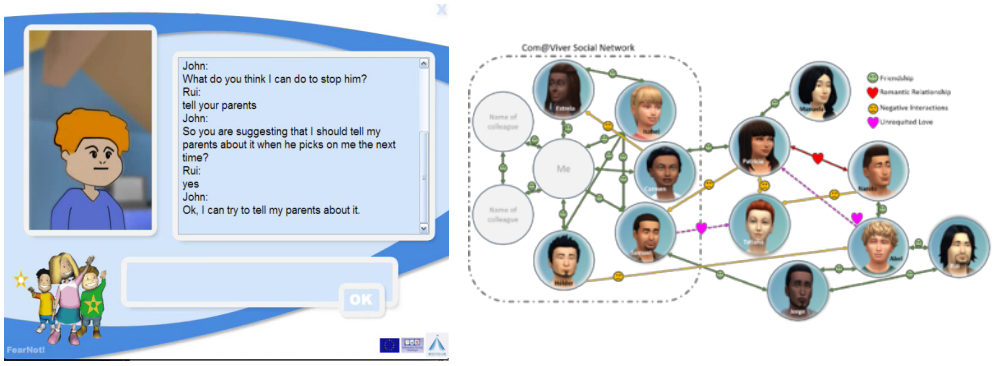


Figure 11.5 Examples of scenarios of use of SIAs to address problems of bullying and cyberbullying (a) FearNot! and (b) Conviver.

to recognize or modulate the user's emotional state, at a given time, or as a response to a given situation; and to be able to communicate effectively in response [Paiva et al. 2017]. This effective communication might include both the agent's ability to convey its emotions through some "embodiment". In addition, some studies have suggested that the embodiment or appearance of the socially interactive agent can have affect on how users respond to it. However, SIAs can be embodied in many different ways. They can be portrayed as a 3D virtual character in a virtual world; as a 2D character on a screen; as a conversational system such as Alexa; disembodied like Cortana or Siri; or even physically embodied as a very realistic social robot like Erica [Glas et al. 2016]. This wide variety of embodiment possibilities leads us to question if the degree of the embodiment may act just as a mere facilitator of the social interaction, or have some impact on the empathic responses as well as prosocial actions by people interacting with them. In fact, we may question if a physical body is better than a virtual, or no body at all for the SIA. In a study by Seo et. al. [Seo et al. 2015] empathy responses to a physical or a virtual "robot" were compared. The main question addressed was: how do people empathize with a physical or a virtual (simulated) robot when something bad happens to it. The results reported suggest that people may empathize more with a physical robot than a virtual one. Indeed it has been shown that empathy display may lead to improved interpersonal relations, with users who consider an empathic robot more as a friend in comparison to a robot not displaying that feature [Pereira et al. 2010]. In a recent study comparing embodied agents (a robot) versus disembodied ones, people interacted with a prosocial agent and a selfish agent in a variant of a public goods game [Correia et al. 2020] (see picture 11.6 for the illustration of the robotic embodiment in this study). The study showed that when the agents were "disembodied", prosocial agents were rated more positively and selfish agents rated more negatively, which is what one would expect. However,

Figure 11.6 Example of robotic embodiment: EMYS from [Correia et al. 2020].

when agents were ”embodied” this effect did not occur, which means that although the social aspects achieved by embodiment can positively affect the emotional responses to agents (as is usually the case), the ”embodiment” itself may mask selfish behaviours from the agents. That is, embodied affordances of the agents seem to lead people to consider additional aspects during the interaction, and the behaviour itself, such as acting selfishly, becomes less salient when compared with other features associated with embodiment. Thus, the ”type” as well as the existence of embodiment in SIAs matter in prosocial contexts [Correia et al. 2020]. Surprisingly, robots that display lower levels of human resemblance seem to be more effective at triggering prosocial behaviours from their human users [De Kleijn et al. 2019]. The study reported in [De Kleijn et al. 2019] suggests that, although appearance might have an effect on fairness, it nevertheless fails to affect prosocial behaviour. Other studies looking at prosocial behaviour in the context of HRI, have used a variety of different robots, making their results hard to compare and leaving the issue of embodiment still largely unresolved. A few authors have, however, already developed social robots especially for this purpose [Sarabia et al. 2013] showing Dona [Kim et al. 2010]), a robot developed for the purpose of collecting money from kind passersby to donate it to charity.

Designing social agents that can successfully exhibit and evoke empathy (and the resulting prosocial outcomes) requires paying special attention to various factors: such as the characteristics of the agent (e.g., its embodiment, or physical appearance- see Chapter 4 of this book [McDonnell and Mutlu 2020]), the dialogue that the agent is able to establish, the social and emotional responses, the non-verbal behaviours, the characteristics of the user, the details of the situation and the mechanisms and modulation processes that can affect the empathic response (e.g. degree of familiarity with the agent, signaling of the need for help, prior social relation with the users) [Paiva et al. 2017].

In fact, emotions have been shown to positively impact prosocial behaviour towards social agents, including context-based amalgamations of positive and negative emotions, such as gratitude (expressed when both the human and the social agent cooperate), shame or guilt

(expressed when the social agent defects) or anger (expressed when the human agent defects) (e.g., [De Melo et al. 2009, 2010]). In particular, the display of context-based emotions aligned with prosocial motivations by a social agent, *seems* to have a positive role in prosociality by increasing the participant's level of trust in the agent [Riegelsberger et al. 2003] and perceived likability [Straßmann et al. 2018], although more research on these mechanisms is needed before reaching a final verdict. Research comparing the effect of context-based emotions in prosociality and cooperation between virtual agents controlled by humans (i.e. avatars) versus agents controlled by algorithms yields similar results according to the type of emotion displayed, with the expression of cooperative intentions by the agents having a superior effect in cooperative behaviour towards avatars (compared to virtual agents), but with signaled competition intentions resulting in similar competitive behaviour towards both avatars and virtual agents [de Melo et al. 2018].

In social robotics, the display or priming of emotions by a social robot also seems to be an important factor to determine both how the robot is perceived and how users respond to it. For example, when a robot starts off the conversation by making a remark related to emotions (rather than related to an object), users are more likely to follow its directions and answer its requests [Imai and Narumi 2004]. In addition, the display of emotion can also modulate the effect of empathy on prosocial behaviour. For example, in a study by Kim and colleagues [Kim et al. 2009], a display of negative emotions by a robot, after receiving a penalty for failing a task, can result in empathy towards the robot, with some participants choosing to suffer the penalty in place of the robot. Having a robot displaying empathy or concern for others can also have a positive effect on the users' intention to engage in prosocial behaviour, as demonstrated in a study conducted by Hayes and colleagues, in which the robot either displayed concern for itself or its' programmer while petitioning the human participant to sacrifice his/hers performance in a competitive task [Hayes et al. 2014]. The authors observed that participants were more likely to help the robot when it displayed concern for others than when it was egotistically motivated and that the level of empathy felt towards the robot was a predictor of the users' likelihood of offering assistance to the robot.

One of the important factors that influence prosociality is the agent's own behaviour during the interaction. This is particularly relevant in scenarios with social dilemmas where the interactions seem to require a certain level of reciprocity or interactive consideration of the other player's strategy [Straßmann et al. 2018]. However, the behaviour of an agent after a prosocial (or antisocial) decision can also influence the user's subsequent actions towards the agent. For example, when the agent does not cooperate, studies have found that negative responses (decreased trust) from the user can be diminished when the virtual agent blushes [Dijk et al. 2011]. Similarly, the non-verbal behaviour adopted by the robot can also help the SIA to evoke prosocial behaviour. For example, one study demonstrated that receiving a reciprocal hug from a robot might lead individuals to donate more money to charity [Shiomi

et al. 2017]. Another study, found a tendency for participants to help a robot complete a task more often after the robot introduced itself with a handshake [Avelino et al. 2018].

Some research also suggests that the goal-orientation manifested by the social agent (i.e. cooperative vs. selfish) can have a positive effect on the participant's own goal orientation, in the context social dilemmas [de Melo et al. 2013, Kulms et al. 2014]. In particular, people are more likely to exhibit cooperative behaviour, guided by the collective profit, when the virtual agent displays similar behaviour, whether that expression is objective (e.g., demonstrated or stated through verbal utterances) or subjective (e.g., demonstrated by the agent's behaviour [de Melo et al. 2013]), which falls in line with social psychology predictions about the role of social values orientations (for a review, see [Bogaert et al. 2008]). Similarly, agents who display cooperative or prosocial emotions are also more likely to evoke prosocial behaviour in social dilemmas, in some cases, regardless of the actual game strategy employed by the agent [De Melo et al. 2010]. Some authors suggest that the interdependence of agents' roles (in this case human and SIA) should also be taken into consideration, however research in this area is still insufficient to draw conclusion [Vásquez and Weretka 2019].

11.5 Towards prosociality in populations with Socially Interactive Agents

So far, we have read that specific agent architectures can be handily implemented to create SIAs that interact with humans, through empathic processing and providing empathic responses. Those goals can be achieved, respectively, through empathy mechanisms, modulation and responses (Sect. 11.3). We have also pointed out that SIAs can trigger altruism in social contexts, which is evidenced by experiments resorting to social dilemmas and economic games involving humans and agents in both physical and virtual environments (Sect. 11.4). Social agents' embodiment, empathy, personality and emotion expression were pointed to positively impact prosociality. Beyond single and short-term interactions (see Chapter 19 on long-term interactions [Kory-Westlund et al. 2020]), a question, however, remains: how can empathic SIAs, embedded in dynamic populations of humans and agents, be used to trigger and stabilize long-term prosociality? In this section we build on the works pointed previously to speculate about the role of SIAs in sustaining prosocial populations of humans and social agents [Paiva et al. 2018]. As defined above, prosocial behaviour can be defined as behaviours that intend to benefit one or more people other than the self. Here, we will mainly focus on prosocial behaviour that involves a cost to the actor – that is, altruistic cooperation – as these acts are particularly hard to trigger. Hopefully, reducing the costs involved in altruism will further facilitate prosocial action. We will discuss SIAs through the lens of evolutionary game theory [Weibull 1997], emphasizing populations and whether certain behaviour can evolve and become evolutionary stable. Within that framework, we will reason about how SIAs can be used to operationalize several cooperation mechanisms [Nowak 2006] pointed out to guarantee the stability of altruistic strategies in social dilemmas.

We envision scenarios in which SIAs can work as instruments to leverage long-term prosociality. To do that, we will discuss the behaviours (strategies) of the agents and discuss four possible classes of agents, with increasing complexity and given the function that they may play in social interactions: 1) *resilient agents*; 2) *reciprocal agents*; 3) *information-sharing agents* and 4) *emotional-signaling agents*. As will be noted, research on SIAs (done in the past 20 years and hopefully in the future) plays a vital role in designing all classes of agents encompassing empathy mechanisms, modulation and responses, and supported by tools such as automated learning, planning, verbal communication, emotional expression and emotional recognition. All these domains are likely to fundamentally impact the design of hybrid populations of (prosocial) socially interactive agents.

11.5.1 Resilient agents

One of the positive effects of SIAs can simply accrue from revealing a fixed prosocial behaviour over time. A agent that acts systematically the same way showing the others its prosocial behaviour. We call these resilient agents. These are, naturally, some of the simplest agents one can think of. In fact, no sophisticated agent architecture is needed to generate such type of behaviour. However, in the context of a population, having a fixed behaviour can affect the overall population dynamics in at least two ways. First, a small fraction of prosocial agents may suffice to reach a critical mass of cooperators above which a population can self-organize towards full cooperation. Second, the existence of agents reveling a fixed prosocial behaviour can incentive others to follow a similar strategy, thus triggering cascades of cooperation through conformist learning or social contagion.

Regarding the first point, we shall refer some previous works that show, precisely, how fixed behaviour can trigger long-term prosociality in a population. Pacheco et al. showed that, in a population of adaptive agents, the existence of a small fraction of obstinate cooperators – defined as those who never change their behaviour over time – is able to change the evolutionary dynamics of a population towards the co-existence of a majority of cooperators [Pacheco and Santos 2011]. The igniting effect of resilient cooperators can be extended to interactions that entail coordination dynamics. Take the example of situations in which a minimal fraction of cooperators is required to achieve a collective goal – a dilemma said to capture the perils of climate change negotiations or simpler mundane tasks such as taking part in a band or team project [Santos et al. 2020]. In those dilemmas, a minimal fraction of cooperators may facilitate collective success and, as such, provide extra incentives for cooperation. Resilient agents may contribute to reaching such thresholds. In the flavour of simple agents contributing to potentiate human coordination, Shirado et al. showed that simple artificial agents with random behaviour, placed in central locations of a social network, can facilitate coordination in human groups [Shirado and Christakis 2017]. Several contexts where cooperation requires extra incentives, however, may not configure the coordination dilemma that resilient unconditional

agents may be suitable to solve. More complex interaction paradigms require more complex agents that make better use of the current capabilities of the SIAs (see below).

Regarding the potential effects of resilient agents through social contagion, we shall refer that imitation was suggested as a relevant enabler of cooperation evolution in humans, leading to cascades of cooperation [Fowler and Christakis 2010]. Experiments by Fowler et al. show that individuals who cooperate tend to influence positively the cooperation level of individuals up to three degrees of separation – that is, a cooperator contributes to increasing the chances that a (1) direct friend, (2) a friend of that friend and (3) a friend of a friend of a friend also cooperate. This reveals the potential overreaching effect that an agent with a prosocial behaviour can have in a networked population. In general, the prospective benefits of resilient prosocial agents is highlighted by works showing that, in human social networks, altruist individuals tend to be connected with altruist neighbors [Leider et al. 2009]. We shall also mention that conformism – i.e., adopting the most common behaviour in a population – is a form of learning also pointed as fundamental in the evolution of cooperation in human societies [Guzmán et al. 2007]. In situations where humans resort to conformism to adapt their behaviour, yet again, observing prosocial agents may increase the chances of behaving altruistically, due to the increase of prosocial models to conform with. As far as we know, it remains an open question knowing whether contagious or conformist cooperation from virtual or robotic agents to humans has similar characteristics as those observed in human social network – e.g., three degrees of separation in positive influence – and which social capabilities are required by the former for that purpose.

One may question however, if these agents are actually SIAs, or agents at all. We however believe that in this context, the term agent and SIA can and should be used to represent the automatic artificial entities that will exist in a society, allowing us to simulate the effects of different behaviours and thus analyse at a macro-level the emergence of prosociality in hybrid societies of humans and technology.

11.5.2 Reciprocal agents

Reciprocal agents introduce a layer of complexity when compared with resilient (unconditional) agents. These agents have memory, are able to recognize their peers' strategies and respond accordingly. Reciprocity (namely direct reciprocity) is known as an important cooperation mechanism [Nowak 2006]. Tit-for-Tat (TFT) is a prototypical example of strategy that can be used by these agents and sustain high levels of cooperation, as identified by Axelrod and Rapoport in the 80s [Axelrod and Hamilton 1981]. This strategy postulates that, in the context of repeated altruistic interactions, individuals should start by cooperating and defect after an opponent defects. If a significant number of TFT agents are introduced in a population, cooperation is able to be stabilized [Imhof et al. 2005]. As a result, a certain fraction of artificial agents, with a judicious choice of reciprocal behaviour, may render cooperation a stable strategy in a population of humans and agents. More recently, Mao et. showed that, in

fact, a small fraction of reciprocal agents with a resilient behaviour – that cooperate until an opponent defects, always defecting afterwards, a strategy also coined Grimm Trigger – is able to significantly increase cooperation levels in a population at large [Mao et al. 2017].

Interactions in the real-world are not constrained to pairwise interactions, where agents decide to cooperate or not with a single opponent. In the context of multiplayer interactions, the decision-making principles encapsulated in TFT can be extended to account for information about a distribution of strategies in a group [Hilbe et al. 2017, Pinheiro et al. 2014]. Also, in scenarios where a critical number of pro-social agents is needed for a collective goal to be achieved, reciprocal agents can be employed to sustain cooperation. In this context, reciprocal agents can use information about their own strategy, on top of information about opponents' previous strategies. Recent work shows that high levels of cooperation and group success can be achieved if agents reciprocate based on their success history and on the strategies anticipated to be played by the others in a group [Santos et al. 2020]. In the context of multiplayer ultimatum games, it was also shown that a small fraction resilient prosocial agents – that give up their payoff to sustain fair outcomes when sharing a given resource – can significantly alter the dynamics in a population of adaptive agents, such that prosocial strategies become stable and prevalent in the long-run [Santos et al. 2019].

We foresee research on empathy and prosociality in SIA being employed in the context of reciprocal agents along, at least, two lines: First of all, empathy mechanisms – as introduced in Sect. 11.3 – are required so that agents are able to recognize others' emotions, goals and intentions. As we have just read, anticipating the intentions of agents in the context of cooperation dilemmas is central to devise conditional strategies that effectively support prosociality. Recent works stress that individuals' behaviour can be anticipated through non-verbal expressions, which allows anticipating humans' reaction to negotiation offers [Park et al. 2013]. Naturally, besides being able to recognize the prosocial intentions of their opponents, SIA can use empathic response channels (also alluded to in Sect. 11.3) to convey their intention to humans, so that the latter can themselves reciprocate. The usage of empathy mechanisms and empathy responses is particularly important in situations where information about the past behaviour of individuals cannot be directly accessed, either because information is not accessible or reliable, or because individuals are interacting for the first time with a specific SIA. Second, research on SIAs, again in the area of empathy mechanisms and user modelling, can prove fundamental in developing agents that avoid getting stuck in long defection periods, after erroneous moves by humans or other agents. One well-known drawback of TFT, when a population at large is using it, is the inability to recover from errors when an isolated defection move (possibly done by mistake of misinterpreted) is done. If this occurs, an opponent using TFT will also defect, which will lead to a subsequent wave of defections. Alternative strategies, such as the Win Stay Lose Shift [Nowak and Sigmund 1993] or Tit-for-two-Tats (TF2T) [Axelrod and Hamilton 1981] were proposed, precisely, to help solving this drawback – introducing others, such as, in the case of TF2T, being prone to

exploitability by more aggressive strategies. In this realm, SIAs can be used to devise agents that successfully convey their real intentions, thus avoiding incurring in loops of defection after a mistake. Likewise, a SIA can be used to recognize errors by other humans or agents, thus being forgiving while remaining non-exploitable. In this context, research on having artificial agents justifying their erroneous moves [Correia et al. 2018a] may provide important advances.

11.5.3 Information-sharing agents

Increasing, once again, the complexity of the considered agents, we foresee the potential benefits of employing SIAs to interact with humans being able to keep memory about others in a population, recognize interactions and share information about interacting individuals. We call these information-sharing agents. The information obtained and shared can be the result of internal mechanisms as discussed previously. In the context of hybrid populations of humans and artificial agents, such SIAs can handily be used in the context of reputations systems [Resnick et al. 2000, Sabater and Sierra 2005] and indirect reciprocity [Nowak and Sigmund 1998], particularly in situations where it is costly for humans to share such information about others [Santos et al. 2018a]. Systems of indirect reciprocity occur when individuals discriminate their actions based on what peers did to others, in the past. The simplest indirect reciprocity systems, build upon the concept of image score [Nowak and Sigmund 1998], occur when individuals cooperate, consequently gaining a positive reputation. Others will then use that reputational uplift to cooperate back. We shall note that these simple reputations systems support a myriad of e-commerce and economy sharing platforms. More complex indirect reciprocity systems – in which altruistic cooperation can become prevalent over time – consider that, for example, reputations are attributed based on the reputation of the individual cooperating and the reputation historic of the individual being helped [Santos et al. 2018b]. In this context, a SIA would need to keep a record of the interacting individuals' reputations, identify the valence of the employed action, and attribute a new reputation based on a given reputation update rule (for example, a rule stating that if an individual with a bad reputation helps an opponent with a good reputation, the helping individual deserves to recover a good reputation). After this process, agents would need to share the new information about the observed individuals, to other SIAs or, potentially, humans. If information about interacting individuals is not readily available or can only be accessed with a high degree of noise, SIAs might be called to identify the intentions of individuals through emotion recognition technologies. Likewise, the reputations of interacting individuals might be communicated through numerical scores, natural language, but also emotion expression (e.g., revealing an angry face whenever an individual deserves a bad reputation).

An insightful connection between empathy and the evolution of cooperation was recently suggested, again in the context of indirect reciprocity systems [Radzvilavicius et al. 2019]. In these systems, one rule to attribute reputations – named stern-judging [Pacheco et al. 2006]

– was shown to combine simplicity with high cooperation levels, guaranteeing the stability of altruism in populations with different sizes and composed by agents with simple cognitive abilities [Santos et al. 2018b]. Stern-judging states that agents should be considered *good when they cooperate with a good opponent or defect with a bad opponent; all else being considered bad*. While this norm promotes high levels of cooperation whenever reputations are able to spread fast in a population and become public (e.g., though gossip), there are some drawbacks when considering private reputations. In particular, two individuals may disagree on how they regard a third peer, and these incongruities can hamper cooperation when stern-judging is the reputation rule prevailing in a population [Hilbe et al. 2018]. Radzvilavicius et al. showed that, in this case, cooperation requires empathic individuals. In [Radzvilavicius et al. 2019], the authors suggest that, when one individual A is judging the behaviour of an individual B (after B plays against a third individual, C) then A can use information in an empathic or egocentric fashion: agent A will be empathic when she places herself in the position of B, taking into account the intentions of B in order to judge her behaviour; in this regard, even if A and B have a different opinion over C, A will use the information that B had. On the other hand, agent A will be egocentric when judging B without considering that B can potentially have a different opinion on C (differing from the opinion of A). Radzvilavicius et al. show, mathematically, how empathy can open new routes for the stability of prosociality, in populations of adaptive agents.

Yet again, tools developed to build SIAs, specifically concerning empathy mechanisms, can be handily used to create agents that judge others and share information in an empathic way at a large scale, thus rendering altruism stable in a population. Particularly, an open question in this context relates to understanding which mechanisms enable individuals to know how another agents' reputation is perceived by others [Masuda and Santos 2019]. We believe that solutions for this challenge may be inspired by emotional expression and communication that characterizes SIAs. On the other hand, many of the techniques and tools used to model large populations of disembodied agents can also influence the design and creation of SIAs.

11.5.4 Emotion-signalling agents

Finally, and for the purpose of simulating these societies of agents and humans, we foresee potential benefits in designing emotional or social-signalling agents. These are agents that, on top of having the ability to discriminate based on pre-play signals of their opponents, are themselves able to communicate their intention before (and after) an interaction, resorting to social signals such as emotional expression.

First, let us introduce the relation between signalling, economic games and (altruistic) prosociality. In coordination games with multiple equilibria, arbitrary signals can disrupt the equilibrium of payoff inferior strategies – that is, strategies that constitute a stable equilibrium yet lead to lower payoffs than other stable strategies [Robson 1990]. Let us say that strategy A is a payoff inferior strategy and strategy B is a payoff superior strategy. In a population

fully composed by individuals playing strategy A or strategy B, no mutant strategy can invade and fixate (respectively, mutant strategy B or mutant strategy A). This means that strategy A can be stable despite the fact that it leads to lower payoffs than B. The stability of strategy A can however be disrupted by arbitrary signals. This can occur through so-called secret handshakes: we can conceive a third strategy, C, that develops a signal only recognized by other individuals adopting C; this strategy will behave as strategy B whenever encountering someone also signalling and will behave as strategy A otherwise. C will invade a population fully composed by the (payoff inferior) strategy A, thus leading to payoff superior outcomes.

Signaling suggests an idyllic scenario in coordination games. One can then speculate if the same mechanisms can be used in cooperation scenarios, assuming that cooperators can signal their cooperative intentions (for example by smiling) and only cooperate with those that also signal. In the case of altruistic cooperation – which leads to prisoner's dilemma type of interactions – there is a catch, however: defectors, i.e. those refusing to take prosocial (altruist) actions, can learn how to fake the signals sustaining cooperation. A population fully composed by cooperators that emit an arbitrary signal before playing, and only cooperate with those using the same signal, can be easily exploited by defectors that use the same signal as cooperators. The interrelation of signaling and cooperation is for long known [Robson 1990]. More recent models, however, show that even if cooperation cannot be stabilized through signaling, the possibility that multiple signals can be used allows cooperation to still become prevalent over time [Santos et al. 2011]. The more signals available, the better, and SIAs are indeed agents that use social signals to communicate.

The advantages and limitations of signaling in sustaining cooperation within populations can again illuminate, in our opinion, some future applications of SIAs. As mentioned before, emotion expression (see Sect. 11.3.4) can be conceived a sophisticated form of pre-play signaling. Thus, on the one hand, emotion expression can disrupt defective equilibria and trigger the evolution of prosocial (altruist) actions. On the other hand, novel emotion recognition tools can allow the implementation of improved ways of anticipating the trustworthiness of expressed signals [Lucas et al. 2016], which can contribute to alleviate the biggest peril of signaling: the possibility that malicious agents fake signals and exploit cooperators.

11.5.5 Environment and networks

So far, we discussed the potential role of SIAs at different levels without placing too much emphasis on environment characteristics where human-agent interactions take place. In Section 11.4 we discussed the types of scenarios that these agents can be used. At the level of populations, we should further consider one particular set of important characteristics that are related to the network topology where individuals interact. At the population level, people to not interact with everyone. They are arranged in networks. Some networks, particularly those where individuals are highly heterogeneous in what concerns the number of contacts they have, were shown to facilitate the evolution of cooperation [Santos and Pacheco 2005]. The

node diversity implied by these networks opens up the possibility of thinking, no only about designing SIAs taking that into account, but also about where to place them in a network. In particular, if those networks in the future can have both SIAs and humans in a hybrid population. Future approaches may combine particular SIAs architectures with knowledge about centrality measures of the network positions where those networks should be deployed.

11.6 Summary and Current Challenges

In this chapter we explored how to create socially interactive agents (SIAs) that not only exhibit empathy, but evoke empathy from others, and, as a consequence, foster prosocial behaviour. Before investigating the approaches to build such agents we started by clarifying some of the major concepts in empathic and prosocial agents, having established a framework for thinking, studying and engineering these agents. This framework allows reasoning about the key elements that agents should include to behave prosocially, or trigger prosociality in interaction groups or populations at large. Then, we analysed SIAs at three different levels. At the micro-level, we discussed the computational mechanisms that are needed to build empathic and prosocial SIAs. we review models and architectures that agents can include to be prosocial, and we elaborate on the different approaches that have been taken by researchers in the field. Then we delved more deeply into how empathic and prosocial SIAs interact with humans and the effects that such type of interactions have. Finally, we investigated how these SIAs can be embedded into a large society and explored empathy and prosociality at that macro, societal level. We believe that this last step is critical to bring humans and agents together in large hybrid groups, and that studying how empathy and prosociality in such agents will allow us to face particular societal challenges, such as inequality, tribalism and sustainability.

Naturally, several challenges lie ahead, in the route to 1) design, 2) deploy and 3) evaluate SIAs that promote prosociality in the real world.

There are natural challenges associated with the deployment of such technologies. Foremost, individuals may have concerns about being influenced by technological artefacts. One may as well remember the Facebook emotional contagion experiment [Kramer et al. 2014] and the ethical debates it prompted [Fiske and Hauser 2014, Verma 2014]. In fact, nudging individuals towards more prosocial behaviours should be done considering high standards of transparency and privacy concerns. Experiments of such kind should be conducted according to the close guidance of Institutional Review Boards, and users should be allowed to opt-out from using the suggested technologies. One should also be clear about how SIAs, intended to promote prosociality, can be tuned to trigger harmful behaviours; in that case, this technology should allow for mechanisms that limit its influence.

At the same time, there are several challenges related to designing and deploying SIAs to trigger prosociality in specific contexts. While here we discuss the potential role of SIAs in generic situations, we foresee that specific scenarios may call for specific details in SIAs to

be tuned. The diversity of situations, specially real-world problems, where we envision that prosocial SIA can be applied to is fascinating. We can think of agents having a real impact in our society ranging from sustaining environmentally friendly behaviours, that support diversity and helping behaviours towards out-group members, or that promote sharing actions that mitigate the effects of inequality. As an extra example, SIAs can be used to nudge individuals adopting responsible behaviours in the midst of our current COVID pandemic. Wearing a mask, keeping social distance, refraining from hoarding, avoiding tempting yet crowded places, help vulnerable people in risk groups, are all behaviours that, even if presenting a small cost to oneself, are a necessary step to achieve collective success. SIAs can be used to highlight how such behaviours can become effective (stressing the collective benefits they lead to) or highlight how others may benefit from them (nurturing empathic concerns in users). In these particular scenarios, one can foresee immense challenges: To start with, how to incentivize individuals to interact, in the first place, with such SIA? Second, how to design and deploy SIA in a timely fashion, that makes this technology useful while, at the same time, guaranteeing that the right amount of effort was placed to design effective agents? How to know which users should targeted first, in order to achieve fast and actual beneficial outcomes – again, in a situation so time-sensitive as an ongoing pandemics? How to incentivise mass usage while making sure that the privacy of each individual is being protected? How to deploy effective SIAs that comply with international regulations on data protection? These are natural societal challenges -— beyond the technical ones —- that may lie ahead when deploying prosocial SIAs.

For centuries, the investigation into human nature has tried to answer whether humans are primarily good or bad. Fortunately, despite human nature being guided mostly by self-serving motivations, it is also known that we help each other at our own cost [Paiva et al. 2018]. Empathic and prosocial SIAs can leverage on this characteristic of human nature to foster prosociality in groups and societies, thus contributing to the establishment of the area of Prosocial Computing [Paiva et al. 2018].

Acknowledgements

This work was supported by FCT scholarships (SFRH/BD/118031/2016 and PD/BD/150570/2020), the AGENTS Project (CMU/TIC/0055/2019), TAILOR project (H2020-ICT-48-2020/952215), and HumaneAI-Net Project (H2020-ICT-48-2020/952026). Fernando Santos acknowledges support from the James S. McDonnell Foundation Postdoctoral Fellowship Award. Ana Paiva is the Katherine Hampson Bessell Fellow of the Radcliffe Institute for Advanced Study at Harvard University, and has been partially funded by the fellowship program.

Bibliography

P. Alves-Oliveira, P. Sequeira, F. S. Melo, G. Castellano, and A. Paiva. 2019. Empathic robot for group learning: A field study. *ACM Transactions on Human-Robot Interaction (THRI)*, 8(1): 1–34.

C. A. Anderson and B. J. Bushman. 2002. Human aggression. *Annual review of psychology*, 53.

K. Anderson, E. André, T. Baur, S. Bernardini, M. Chollet, E. Chryssafidou, I. Damian, C. Ennis, A. Egges, P. Gebhard, et al. 2013. The tardis framework: intelligent virtual agents for social coaching in job interviews. In *International Conference on Advances in Computer Entertainment Technology*, pp. 476–491. Springer.

M. Asada. 2015. Towards artificial empathy. *International Journal of Social Robotics*, 7(1): 19–33.

J. Avelino, F. Correia, J. Catarino, P. Ribeiro, P. Moreno, A. Bernardino, and A. Paiva. 2018. The power of a hand-shake in human-robot interactions. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1864–1869. IEEE.

R. Axelrod and W. D. Hamilton. 1981. The evolution of cooperation. *Science*, 211(4489): 1390–1396.

R. S. Aylett, S. Louchart, J. Dias, A. Paiva, and M. Vala. 2005. Fearnot!–an experiment in emergent narrative. In *International Workshop on Intelligent Virtual Agents*, pp. 305–316. Springer.

C. P. Barlett and C. A. Anderson. 2012. Examining media effects: The general aggression and general learning models. *The international encyclopedia of media studies*.

C. D. Batson. 2011. *Altruism in humans*. Oxford University Press, USA.

C. D. Batson. 2014. *The altruism question: Toward a social-psychological answer*. Psychology Press.

C. D. Batson and N. Ahmad. 2001. Empathy-induced altruism in a prisoner's dilemma ii: What if the target of empathy has defected? *European Journal of Social Psychology*, 31(1): 25–36.

C. D. Batson, J. G. Batson, J. K. Slingsby, K. L. Harrell, H. M. Peekna, and R. M. Todd. 1991. Empathic joy and the empathy-altruism hypothesis. *Journal of personality and social psychology*, 61(3): 413.

C. D. Batson, J. G. Batson, R. M. Todd, B. H. Brummett, L. L. Shaw, and C. M. Aldeguer. 1995. Empathy and the collective good: Caring for one of the others in a social dilemma. *Journal of personality and social psychology*, 68(4): 619.

C. D. Batson, D. A. Lishner, and E. L. Stocks. 2015. 13 the empathy–altruism hypothesis. *The Oxford handbook of prosocial behavior*, pp. 259–268.

D. J. Baumann, R. B. Cialdini, and D. T. Kendrick. 1981. Altruism as hedonism: Helping and self-gratification as equivalent responses. *Journal of Personality and Social Psychology*, 40(6): 1039.

C. Becker, H. Prendinger, M. Ishizuka, and I. Wachsmuth. 2005. Evaluating affective feedback of the 3d agent max in a competitive cards game. In *International Conference on Affective Computing and Intelligent Interaction*, pp. 466–473. Springer.

J. Berg, J. Dickhaut, and K. McCabe. 1995. Trust, reciprocity, and social history. *Games and economic behavior*, 10(1): 122–142.

T. W. Bickmore and R. W. Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2): 293–327.

P. Bloom. 2017. *Against empathy: The case for rational compassion*. Random House.

S. Bogaert, C. Boone, and C. Declerck. 2008. Social value orientation and cooperation in social dilemmas: A review and conceptual model. *British Journal of Social Psychology*, 47(3): 453–480.

F. Borgonovi. 2008. Doing well by doing good. the relationship between formal volunteering and self-reported health and happiness. *Social science & medicine*, 66(11): 2321–2334.

H. Boukricha, I. Wachsmuth, M. N. Carminati, and P. Knoeferle. 2013. A computational model of empathy: Empirical evaluation. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 1–6. IEEE.

S. Brave, C. Nass, and K. Hutchinson. 2005. Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International journal of human-computer studies*, 62(2): 161–178.

J. Broekens. 2020. Emotion. In B. Lugrin, C. Pelachaud, and D. Traum, eds., *Handbook on Socially Interactive Agents*. ACM.

J. L. Brown. 1978. Avian communal breeding systems. *Annual Review of Ecology and Systematics*, 9(1): 123–155.

S. L. Brown, R. M. Nesse, A. D. Vinokur, and D. M. Smith. 2003. Providing social support may be more beneficial than receiving it: Results from a prospective study of mortality. *Psychological science*, 14(4): 320–327.

K. E. Buckley and C. A. Anderson. 2006. A theoretical model of the effects and consequences of playing video games. *Playing video games: Motives, responses, and consequences*, pp. 363–378.

P. Campos-Mercade, A. Meier, F. Schneider, and E. Wengström. 2020. Prosociality predicts health behaviors during the covid-19 pandemic. *University of Zurich, Department of Economics, Working Paper*, (346).

C. S. Carter, I. B.-A. Bartal, and E. C. Porges. 2017. The roots of compassion: An evolutionary and neurobiological perspective. *The Oxford handbook of compassion science*, p. 173.

G. Castellano, A. Paiva, A. Kappas, R. Aylett, H. Hastie, W. Barendregt, F. Nabais, and S. Bull. 2013. Towards empathic virtual and robotic tutors. In *International conference on artificial intelligence in education*, pp. 733–736. Springer.

F. Correia, C. Guerra, S. Mascarenhas, F. S. Melo, and A. Paiva. 2018a. Exploring the impact of fault justification in human-robot trust. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 507–513. International Foundation for Autonomous Agents and Multiagent Systems.

F. Correia, S. Mascarenhas, R. Prada, F. S. Melo, and A. Paiva. 2018b. Group-based emotions in teams of humans and robots. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*, pp. 261–269.

F. Correia, S. Gomes, S. Mascarenhas, F. S. Melo, and A. Paiva. 2020. The dark side of embodiment teaming up with robots vs disembodied agents. In *Proceedings of the Robotics Science and Systems RSS'2020*.

A. Costantini, A. Scalco, R. Sartori, E. M. Tur, and A. Ceschi. 2019. Theories for computing prosocial behavior. *Nonlinear Dynamics Psychol. Life Sci*, 23: 297–313.

S. M. Coyne, L. M. Padilla-Walker, H. G. Holmgren, E. J. Davis, K. M. Collier, M. K. Memmott-Elison, and A. J. Hawkins. 2018. A meta-analysis of prosocial media on prosocial behavior, aggression, and empathic concern: A multidimensional approach. *Developmental Psychology*, 54(2): 331–347.

J. Crocker, A. Canevello, and A. A. Brown. 2017. Social motivation: Costs and benefits of selfishness and otherishness. *Annual Review of Psychology*, 68: 299–325.

M. H. Davis. 2018. *Empathy: A social psychological approach*. Routledge.

R. M. Dawes. 1980. Social dilemmas. *Annual review of psychology*, 31(1): 169–193.

R. De Kleijn, L. van Es, G. Kachergis, and B. Hommel. 2019. Anthropomorphization of artificial agents leads to fair and strategic, but not altruistic behavior. *International Journal of Human-Computer Studies*, 122: 168–173.

C. de Melo, P. Carnevale, and J. Gratch. 2013. People's biased decisions to trust and cooperate with agents that express emotions. In *Proc. AAMAS*.

C. M. De Melo, L. Zheng, and J. Gratch. 2009. Expression of moral emotions in cooperating agents. In *International Workshop on Intelligent Virtual Agents*, pp. 301–307. Springer.

C. M. De Melo, P. Carnevale, and J. Gratch. 2010. The influence of emotions in embodied agents on human decision-making. In *International Conference on Intelligent Virtual Agents*, pp. 357–370. Springer.

C. M. de Melo, P. Khooshabeh, O. Amir, and J. Gratch. 2018. Shaping cooperation between humans and agents with emotion expressions and framing. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2224–2226. International Foundation for Autonomous Agents and Multiagent Systems.

F. De Vignemont and T. Singer. 2006. The empathic brain: how, when and why? *Trends in cognitive sciences*, 10(10): 435–441.

F. B. De Waal. 2007. The 'russian doll'model of empathy and imitation. *On being moved: From mirror neurons to empathy*, pp. 35–48.

F. B. De Waal. 2008. Putting the altruism back into altruism: the evolution of empathy. *Annu. Rev. Psychol.*, 59: 279–300.

C. Dijk, B. Koenig, T. Ketelaar, and P. J. de Jong. 2011. Saved by the blush: being trusted despite defecting. *Emotion*, 11(2): 313.

E. W. Dunn, L. B. Aknin, and M. I. Norton. 2014. Prosocial spending and happiness: Using money to benefit others pays off. *Current Directions in Psychological Science*, 23(1): 41–47.

N. Eisenberg and T. L. Spinrad. 2014. Multidimensionality of prosocial behavior: Rethinking the conceptualization and development of prosocial behavior. pp. 17–39.

N. Eisenberg, S. K. VanSchyndel, and T. L. Spinrad. 2016. Prosocial motivation: Inferences from an opaque body of work. *Child Development*, 87(6): 1668–1678.

P. Ekman. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60): 16.

E. Fehr and U. Fischbacher. 2003. The nature of human altruism. *Nature*, 425(6960): 785–791.

P. C. Ferreira, A. V. Simão, A. Paiva, and A. Ferreira. 2020a. Responsive bystander behaviour in cyberbullying: a path through self-efficacy. *Behaviour & Information Technology*, 39(5): 511–524.

P. C. Ferreira, A. Simão, A. Paiva, C. Martinho, R. Prada, A. Ferreira, and F. Santos. 2020b. Exploring empathy in cyberbullying with serious games (under review). Technical report, University of Lisbon.

S. T. Fiske and R. M. Hauser, 2014. Protecting human research participants in the age of big data.

J. H. Fowler and N. A. Christakis. 2010. Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences*, 107(12): 5334–5338.

A. D. Galinsky, W. W. Maddux, D. Gilin, and J. B. White. 2008. Why it pays to get inside the head of your opponent: The differential effects of perspective taking and empathy in negotiations. *Psychological science*, 19(4): 378–384.

L. Gamberini, L. Chittaro, A. Spagnolli, and C. Carlesso. 2015. Psychological response to an emergency in virtual reality: Effects of victim ethnicity and emergency type on helping behavior and navigation. *Computers in Human Behavior*, 48: 104–113.

D. A. Gentile, C. A. Anderson, S. Yukawa, N. Ihori, M. Saleem, L. K. Ming, A. Shibuya, A. K. Liau, A. Khoo, B. J. Bushman, et al. 2009. The effects of prosocial video games on prosocial behaviors: International evidence from correlational, longitudinal, and experimental studies. *Personality and Social Psychology Bulletin*, 35(6): 752–763.

D. F. Glas, T. Minato, C. T. Ishi, T. Kawahara, and H. Ishiguro. 2016. Erica: The erato intelligent conversational android. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 22–29. IEEE.

C. T. Gloria and M. A. Steinhardt. 2016. Relationships among positive emotions, coping, resilience and mental health. *Stress and Health*, 32(2): 145–156.

A. P. Goldstein et al. 1994. *The Prosocial Gang: Implementing Aggression Replacement Training.* ERIC.

B. Gonsior, S. Sosnowski, C. Mayer, J. Blume, B. Radig, D. Wollherr, and K. Kühnlenz. 2011. Improving aspects of empathy and subjective performance for hri through mirroring facial expressions. In *2011 RO-MAN*, pp. 350–356. IEEE.

N. M. Gotts, J. G. Polhill, and A. N. R. Law. 2003. Agent-based simulation in the study of social dilemmas. *Artificial Intelligence Review*, 19(1): 3–92.

W. G. Graziano, M. M. Habashi, B. E. Sheese, and R. M. Tobin. 2007. Agreeableness, empathy, and helping: A person× situation perspective. *Journal of personality and social psychology*, 93(4): 583.

T. Greitemeyer and D. O. Mügge. 2014. Video games do affect social outcomes: A meta-analytic review of the effects of violent and prosocial video game play. *Personality and social psychology bulletin*, 40(5): 578–589.

R. A. Guzmán, C. Rodríguez-Sickert, and R. Rowthorn. 2007. When in rome, do as the romans do: the coevolution of altruistic punishment, conformist learning, and cooperation. *Evolution and Human Behavior*, 28(2): 112–117.

M. M. Habashi, W. G. Graziano, and A. E. Hoover. 2016. Searching for the prosocial personality: A big five approach to linking personality and prosocial behavior. *Personality and Social Psychology Bulletin*, 42(9): 1177–1192.

W. D. Hamilton. 1964. The genetical evolution of social behaviour. ii. *Journal of theoretical biology*, 7(1): 17–52.

W. D. Hamilton. 1972. Altruism and related phenomena, mainly in social insects. *Annual Review of Ecology and systematics*, 3(1): 193–232.

B. Hayes, D. Ullman, E. Alexander, C. Bank, and B. Scassellati. 2014. People help robots who help others, not robots who help themselves. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pp. 255–260. IEEE.

C. Hilbe, L. A. Martinez-Vaquero, K. Chatterjee, and M. A. Nowak. 2017. Memory-n strategies of direct reciprocity. *Proceedings of the National Academy of Sciences*, 114(18): 4715–4720.

C. Hilbe, L. Schmid, J. Tkadlec, K. Chatterjee, and M. A. Nowak. 2018. Indirect reciprocity with private, noisy, and incomplete information. *Proceedings of the National Academy of Sciences*, 115(48): 12241–12246.

B. E. Hilbig, A. Glöckner, and I. Zettler. 2014. Personality and prosocial behavior: Linking basic traits and social value orientations. *Journal of Personality and Social Psychology*, 107(3): 529.

M. L. Hoffman. 2001. *Empathy and moral development: Implications for caring and justice*. Cambridge University Press.

E. N. M. Ibrahim and C. S. Ang. 2018. Communicating empathy: Can technology intervention promote pro-social behavior?—review and perspectives. *Advanced Science Letters*, 24(3): 1643–1646.

M. Imai and M. Narumi. 2004. Robot behavior for encouraging immersion in interaction. *Proceedings of Complex Systems Intelligence and Modern Technological Applications (CSIMTA 2004), Cherbourg, France*, pp. 591–598.

L. A. Imhof, D. Fudenberg, and M. A. Nowak. 2005. Evolutionary cycles of cooperation and defection. *Proceedings of the National Academy of Sciences*, 102(31): 10797–10800.

C. E. Izard. 2013. *Human emotions*. Springer Science & Business Media.

D. Keltner, A. Kogan, P. K. Piff, and S. R. Saturn. 2014. The sociocultural appraisals, values, and emotions (save) framework of prosociality: Core processes from gene to meme. *Annual review of psychology*, 65: 425–460.

E. H. Kim, S. S. Kwak, and Y. K. Kwak. 2009. Can robotic emotional expressions induce a human to empathize with a robot? In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 358–362. IEEE.

M. S. Kim, B. K. Cha, D. M. Park, S. M. Lee, S. Kwak, and M. K. Lee. 2010. Dona: Urban donation motivating robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 159–160. IEEE.

T. J. King, I. Warren, and D. Palmer. 2008. Would kitty genovese have been murdered in second life? researching the" bystander effect" using online technologies. In *TASA 2008: Re-imagining sociology: the annual conference of The Australian Sociological Association*, pp. 1–23. University of Melbourne.

S.-C. Kolm. 2008. *Reciprocity: An economics of social relations*. Cambridge University Press.

J. M. Kory-Westlund, C. Breazeal, H. Park, and I. Grover. 2020. Long-term interaction. In B. Lugrin, C. Pelachaud, and D. Traum, eds., *Handbook on Socially Interactive Agents*. ACM.

M. D. Kozlov and M. K. Johansen. 2010. Real behavior in virtual environments: Psychology experiments in a simple virtual-reality paradigm using video games. *Cyberpsychology, behavior,*

and social networking, 13(6): 711–714.

A. D. Kramer, J. E. Guillory, and J. T. Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24): 8788–8790.

N. Kramer and A. Manzeschke. 2020. Social reactions to socially interactive agents and their ethical implications. In B. Lugrin, C. Pelachaud, and D. Traum, eds., *Handbook on Socially Interactive Agents*. ACM.

P. Kulms, S. Kopp, and N. C. Krämer. 2014. Let's be serious and have a laugh: Can humor support cooperation with a virtual agent? In *International Conference on Intelligent Virtual Agents*, pp. 250–259. Springer.

B. Latané and J. M. Darley. 1970. *The unresponsive bystander: Why doesn't he help?* Appleton-Century-Crofts.

S. Leiberg, O. Klimecki, and T. Singer. 2011. Short-term compassion training increases prosocial behavior in a newly developed prosocial game. *PloS one*, 6(3).

S. Leider, M. M. Möbius, T. Rosenblat, and Q.-A. Do. 2009. Directed altruism and enforced reciprocity in social networks. *The Quarterly Journal of Economics*, 124(4): 1815–1851.

I. Leite, A. Pereira, S. Mascarenhas, C. Martinho, R. Prada, and A. Paiva. 2013. The influence of empathy in human–robot relations. *International journal of human-computer studies*, 71(3): 250–260.

I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva. 2014. Empathic robots for long-term interaction. *International Journal of Social Robotics*, 6(3): 329–341.

A. Lim and H. G. Okuno. 2015. A recipe for empathy. *International Journal of Social Robotics*, 7(1): 35–49.

D. Lim and D. DeSteno. 2016. Suffering and compassion: The links among adverse life experiences, empathy, compassion, and prosocial behavior. *Emotion*, 16(2): 175.

C. Lisetti, R. Amini, U. Yasavur, and N. Rishe. 2013. I can help you change! an empathic virtual agent delivers behavior change health interventions. *ACM Transactions on Management Information Systems (TMIS)*, 4(4): 1–28.

G. Lucas, G. Stratou, S. Lieblich, and J. Gratch. 2016. Trust me: multimodal signals of trustworthiness. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 5–12.

E. Lukinova and M. Myagkov. 2016. Impact of short social training on prosocial behaviors: An fmri study. *Frontiers in systems neuroscience*, 10: 60.

M. J. Lupoli, L. Jampol, and C. Oveis. 2017. Lying because we care: Compassion increases prosocial lying. *Journal of Experimental Psychology: General*, 146(7): 1026.

L. K. Ma, R. J. Tunney, and E. Ferguson. 2017. Does gratitude enhance prosociality?: A meta-analytic review. *Psychological Bulletin*, 143(6): 601–635.

H. L. Maibom. 2017. Introduction to philosophy of empathy. *The Routledge Handbook to Philosophy of Empathy, New York: Routledge*, pp. 1–10.

A. Mao, L. Dworkin, S. Suri, and D. J. Watts. 2017. Resilient cooperators stabilize long-run cooperation in the finitely repeated prisoner's dilemma. *Nature communications*, 8(1): 1–10.

F. Martela and R. M. Ryan. 2016. The benefits of benevolence: Basic psychological needs, beneficence, and the enhancement of well-being. *Journal of personality*, 84(6): 750–764.

N. Masuda and F. C. Santos. 2019. A mathematical look at empathy. *eLife*, 8.

R. McDonnell and B. Mutlu. 2020. Appearance. In B. Lugrin, C. Pelachaud, and D. Traum, eds., *Handbook on Socially Interactive Agents*. ACM.

S. W. McQuiggan, J. L. Robison, R. Phillips, and J. C. Lester. 2008. Modeling parallel and reactive empathy in virtual agents: an inductive approach. In *AAMAS (1)*, pp. 167–174. Citeseer.

R. R. Morris, K. Kouddous, R. Kshirsagar, and S. M. Schueller. 2018. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. *Journal of medical Internet research*, 20(6): e10148.

L. D. Nelson and M. I. Norton. 2005. From student to superhero: Situational primes shape future helping. *Journal of experimental social psychology*, 41(4): 423–430.

M. Nowak and K. Sigmund. 1993. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature*, 364(6432): 56–58.

M. A. Nowak. 2006. Five rules for the evolution of cooperation. *science*, 314(5805): 1560–1563.

M. A. Nowak and S. Roch. 2007. Upstream reciprocity and the evolution of gratitude. *Proceedings of the royal society B: Biological Sciences*, 274(1610): 605–610.

M. A. Nowak and K. Sigmund. 1998. Evolution of indirect reciprocity by image scoring. *Nature*, 393(6685): 573–577.

M. Ochs, D. Sadek, and C. Pelachaud. 2012. A formal model of emotions for an empathic rational dialog agent. *Autonomous Agents and Multi-Agent Systems*, 24(3): 410–440.

J. M. Pacheco and F. C. Santos. 2011. The messianic effect of pathological altruism. *Pathological Altruism*, p. 300.

J. M. Pacheco, F. C. Santos, and F. A. C. Chalub. 2006. Stern-judging: A simple, successful norm which promotes cooperation under indirect reciprocity. *PLoS computational biology*, 2(12).

A. Paiva, I. Leite, H. Boukricha, and I. Wachsmuth. 2017. Empathy in virtual agents and robots: a survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(3): 1–40.

A. Paiva, F. P. Santos, and F. C. Santos. 2018. Engineering pro-sociality with autonomous agents. In *Thirty-second AAAI conference on artificial intelligence*.

S. Park, S. Scherer, J. Gratch, P. Carnevale, and L.-P. Morency. 2013. Mutual behaviors during dyadic negotiation: Automatic prediction of respondent reactions. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 423–428. IEEE.

L. A. Penner, J. F. Dovidio, J. A. Piliavin, and D. A. Schroeder. 2005. Prosocial behavior: Multilevel perspectives. *Annu. Rev. Psychol.*, 56: 365–392.

E. Pennisi. 2005. How did cooperative behavior evolve? *Science*, 309(5731): 93–93.

A. Pereira, I. Leite, S. Mascarenhas, C. Martinho, and A. Paiva. 2010. Using empathy to improve human-robot relationships. In *International Conference on Human-Robot Personal Relationship*, pp. 130–138. Springer.

S. Pfattheicher, L. Nockur, R. Böhm, C. Sassenrath, and M. B. Petersen. 2020. The emotional path to action: Empathy promotes physical distancing during the covid-19 pandemic.

R. W. Picard, A. Wexelblat, and C. I. N. I. Clifford I. Nass. 2002. Future interfaces: social and emotional. In *CHI'02 Extended Abstracts on Human Factors in Computing Systems*, pp. 698–699.

F. L. Pinheiro, V. V. Vasconcelos, F. C. Santos, and J. M. Pacheco. 2014. Evolution of all-or-none strategies in repeated public goods dilemmas. *PLoS computational biology*, 10(11).

H. Prendinger and M. Ishizuka. 2005. The empathic companion: A character-based interface that addresses users'affective states. *Applied artificial intelligence*, 19(3-4): 267–285.

S. D. Preston and F. B. De Waal. 2002. Empathy: Its ultimate and proximate bases. *Behavioral and brain sciences*, 25(1): 1–20.

S. Prot, D. A. Gentile, C. A. Anderson, K. Suzuki, E. Swing, K. M. Lim, Y. Horiuchi, M. Jelic, B. Krahé, W. Liuqing, et al. 2014. Long-term relations among prosocial-media use, empathy, and prosocial behavior. *Psychological science*, 25(2): 358–368.

G. R. Pursell, B. Laursen, K. H. Rubin, C. Booth-LaForce, and L. Rose-Krasnor. 2008. Gender differences in patterns of association between prosocial behavior, personality, and externalizing problems. *Journal of Research in Personality*, 42(2): 472–481.

A. L. Radzvilavicius, A. J. Stewart, and J. B. Plotkin. 2019. Evolution of empathetic moral evaluation. *eLife*, 8: e44269.

L. T. Rameson, S. A. Morelli, and M. D. Lieberman. 2012. The neural correlates of empathy: experience, automaticity, and prosocial behavior. *Journal of cognitive neuroscience*, 24(1): 235–245.

D. G. Rand and M. A. Nowak. 2013. Human cooperation. *Trends in cognitive sciences*, 17(8): 413–425.

D. G. Rand, J. D. Greene, and M. A. Nowak. 2012. Spontaneous giving and calculated greed. *Nature*, 489(7416): 427–430.

B. Reeves and C. I. Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge university press.

P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman. 2000. Reputation systems. *Communications of the ACM*, 43(12): 45–48.

J. Riegelsberger, M. A. Sasse, and J. D. McCarthy. 2003. The researcher's dilemma: evaluating trust in computer-mediated communication. *International Journal of Human-Computer Studies*, 58(6): 759–781.

L. D. Riek, P. C. Paul, and P. Robinson. 2010. When my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry. *Journal on Multimodal User Interfaces*, 3(1-2): 99–108.

A. J. Robson. 1990. Efficiency in evolutionary games: Darwin, nash and the secret handshake. *Journal of theoretical Biology*, 144(3): 379–396.

S. H. Rodrigues, S. Mascarenhas, J. Dias, and A. Paiva. 2015. A process model of empathy for virtual agents. *Interacting with Computers*, 27(4): 371–391.

R. S. Rosenberg, S. L. Baughman, and J. N. Bailenson. 2013. Virtual superheroes: Using superpowers in virtual reality to encourage prosocial behavior. *PloS one*, 8(1).

A. C. Rumble, P. A. Van Lange, and C. D. Parks. 2010. The benefits of empathy: When empathy may sustain cooperation in social dilemmas. *European Journal of Social Psychology*, 40(5): 856–866.

J. Sabater and C. Sierra. 2005. Review on computational trust and reputation models. *Artificial intelligence review*, 24(1): 33–60.

J. Sabourin, B. Mott, and J. Lester. 2011. Computational models of affect and empathy for pedagogical virtual agents. In *Standards in emotion modeling, Lorentz Center International Center for workshops in the Sciences*. Citeseer.

M. Saleem, C. A. Anderson, and D. A. Gentile. 2012. Effects of prosocial, neutral, and violent video games on children's helpful and hurtful behaviors. *Aggressive behavior*, 38(4): 281–287.

D. Sally. 1995. Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and society*, 7(1): 58–92.

F. C. Santos and J. M. Pacheco. 2005. Scale-free networks provide a unifying framework for the emergence of cooperation. *Physical Review Letters*, 95(9): 098104.

F. C. Santos, J. M. Pacheco, and B. Skyrms. 2011. Co-evolution of pre-play signaling and cooperation. *Journal of Theoretical Biology*, 274(1): 30–35.

F. P. Santos, J. M. Pacheco, and F. C. Santos. 2018a. Social norms of cooperation with costly reputation building. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

F. P. Santos, F. C. Santos, and J. M. Pacheco. 2018b. Social norm complexity and past reputations in the evolution of cooperation. *Nature*, 555(7695): 242–245.

F. P. Santos, J. M. Pacheco, A. Paiva, and F. C. Santos. 2019. Evolution of collective fairness in hybrid populations of humans and agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6146–6153.

F. P. Santos, S. F. Mascarenhas, F. C. Santos, F. Correia, S. Gomes, and A. Paiva. 2020. Picky losers and carefree winners prevail in collective risk dilemmas with partner selection. *Autonomous Agents and Multi-Agent Systems*, 34(40).

M. Sapouna, D. Wolke, N. Vannini, S. Watson, S. Woods, W. Schneider, S. Enz, L. Hall, A. Paiva, E. André, et al. 2010. Virtual learning intervention to reduce bullying victimization in primary school: a controlled trial. *Journal of Child Psychology and Psychiatry*, 51(1): 104–112.

M. Sarabia, T. Le Mau, H. Soh, S. Naruse, C. Poon, Z. Liao, K. C. Tan, Z. J. Lai, and Y. Demiris. 2013. icharibot: Design and field trials of a fundraising robot. In *International Conference on Social Robotics*, pp. 412–421. Springer.

E. G. Schellenberg, K. A. Corrigall, S. P. Dys, and T. Malti. 2015. Group music training and children's prosocial skills. *PLoS One*, 10(10).

K. R. Scherer. 1988. Criteria for emotion-antecedent appraisal: A review. In *Cognitive perspectives on emotion and motivation*, pp. 89–126. Springer.

T. Schramme. 2017. Empathy and altruism. *The Routledge handbook of philosophy of empathy*, pp. 203–214.

S. H. Seo, D. Geiskkovitch, M. Nakane, C. King, and J. E. Young. 2015. Poor thing! would you feel sorry for a simulated robot? a comparison of empathy toward a physical and a simulated robot. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 125–132. IEEE.

P. Sequeira, P. Alves-Oliveira, T. Ribeiro, E. Di Tullio, S. Petisca, F. S. Melo, G. Castellano, and A. Paiva. 2016. Discovering social interaction strategies for robots from restricted-perception wizard-of-oz studies. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 197–204. IEEE.

M. Shiomi, A. Nakata, M. Kanbara, and N. Hagita. 2017. A hug from a robot encourages prosocial behavior. In *2017 26th IEEE international symposium on robot and human interactive communication (RO-MAN)*, pp. 418–423. IEEE.

H. Shirado and N. A. Christakis. 2017. Locally noisy autonomous agents improve global human coordination in network experiments. *Nature*, 545(7654): 370–374.

M. Slater, A. Rovira, R. Southern, D. Swapp, J. J. Zhang, C. Campbell, and M. Levine. 2013. Bystander responses to a violent incident in an immersive virtual environment. *PloS one*, 8(1).

C. Straßmann, A. M. Rosenthal-von der Pütten, and N. C. Krämer. 2018. With or against each other? the influence of a virtual agent's (non) cooperative behavior on user's cooperation behavior in the prisoners' dilemma. *Advances in Human-Computer Interaction*, 2018.

A. Tavoni, A. Dannenberg, G. Kallis, and A. Löschel. 2011. Inequality, communication, and the avoidance of disastrous climate change in a public goods game. *Proceedings of the National Academy of Sciences*, 108(29): 11825–11829.

I. Thielmann, G. Spadaro, and D. Balliet. 2020. Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological Bulletin*, 146(1): 30–90.

M. Tomasello and A. Vaish. 2013. Origins of human cooperation and morality. *Annual review of psychology*, 64: 231–255.

R. L. Trivers. 1971. The evolution of reciprocal altruism. *The Quarterly review of biology*, 46(1): 35–57.

P. A. Van Lange, J. Joireman, C. D. Parks, and E. Van Dijk. 2013. The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes*, 120(2): 125–141.

P. A. Van Lange, D. P. Balliet, C. D. Parks, and M. Van Vugt. 2014. *Social dilemmas: Understanding human cooperation*. Oxford University Press.

J. Vásquez and M. Weretka. 2019. Affective empathy in non-cooperative games. *Games and Economic Behavior*.

I. M. Verma. 2014. Editorial expression of concern: Experimental evidence of massivescale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, p. 201412469.

J. W. Weibull. 1997. *Evolutionary game theory*. MIT press.

S. A. West, A. S. Griffin, and A. Gardner. 2007. Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of evolutionary biology*, 20(2): 415–432.

M. Wooldridge. 2003. *Reasoning about rational agents*. MIT press.

J. Wu, D. Balliet, L. S. Peperkoorn, A. Romano, and P. A. Van Lange. 2020. Cooperation in groups of different sizes: The effects of punishment and reputation-based partner choice. *Frontiers in Psychology*, 10: 2956.

S. X. Xiao, E. C. Hashi, K. M. Korous, and N. Eisenberg. 2019. Gender differences across multiple types of prosocial behavior in adolescence: A meta-analysis of the prosocial tendency measure-revised (ptm-r). *Journal of adolescence*, 77: 41–58.

Ö. Yalçın and S. DiPaola. 2019a. Evaluating levels of emotional contagion with an embodied conversational agent. In *Proceedings of the 41st annual conference of the cognitive science society*.

Ö. N. Yalçın and S. DiPaola. 2018. A computational model of empathy for interactive agents. *Biologically inspired cognitive architectures*, 26: 20–25.

Ö. N. Yalçın and S. DiPaola. 2019b. Modeling empathy: building a link between affective and cognitive processes. *Artificial Intelligence Review*, pp. 1–24.