Evol Biol (2011) 38:124–143
DOI 10.1007/s11692-011-9119-6

SYNTHESIS PAPER

Physical Constraints on the Evolution of Cooperation

Anton J. M. Dijker

Received: 1 December 2010 / Accepted: 11 March 2011 / Published online: 22 March 2011

Abstract The evolution of psychological adaptations for cooperation is still puzzling due to a tendency to frame social interaction in mathematical and game-theoretical terms, without systematically examining its causal structure and underlying mechanisms. Complementarily, empirical approaches to cooperation tend to focus on isolated components of mechanisms without sufficiently indicating how different components are combined into a single mechanism and different mechanisms fit into a single organism. An alternative approach to the evolution of cooperation is proposed, starting from a description of basic physical properties of individuals and their environment, and the limited physical or mechanistic possibilities to generate adaptive responses to those properties. This approach reveals that some forms of symmetrical cooperation do not require mechanisms "specifically designed for" benefiting others, whereas effective helping requires a specific mechanism that relatively unconditionally and persistently responds to the vulnerability of other individuals. Unraveling the causal structure of different types of other-benefiting shows that a mechanism for asymmetrical helping may considerably improve symmetrical cooperation through properties such as tolerance, patience, and the human capacity to experience a wide variety of moral emotions. The proposed mechanistic approach to cooperation provides the mathematical/game-theoretical approach with realistic assumptions about psychological adaptations, and helps to integrate the scattered facts about mechanisms gathered by the empirical approach. It also helps to build bridges between the two approaches by providing a common language for thinking about psychological mechanisms.

Keywords Cooperation · Helping · Psychological mechanisms · Game theory · Evolution · Physical properties

> One of the facts that we must face about evolutionary systems is that their simple organizational principles can imply extraordinarily subtle properties. Indeed, part of the dilemma that many students of mind now face is not that they do not know enough facts on which to base a theory, but rather they do not know which facts are principles and which are epiphenomena, and how to derive the multitudinous consequences that occur when a few principles act together (Grossberg 1980, p. 3).

Introduction

Broadly conceived, cooperation refers to organisms behaving in such a way that they improve each others' fitness or reproductive success. Biologists, psychologists, and economists attempting to identify, describe, and integrate the psychological mechanisms underlying these behaviors, and to explain their evolution, are faced with two major problems. First, these behaviors are extremely divers and may be caused by both common and unique mechanisms which may be difficult to distinguish. Second, it is unclear what realistic assumptions should be made about the general physical or neural properties of the mechanisms one is looking for. This paper argues that both problems are

A. J. M. Dijker (✉)
Faculty of Health, Medicine and Life Sciences,
Department of Health Promotion, Maastricht University,
CAPHRI, PO Box 616, 6200 MD Maastricht, The Netherlands
e-mail: a.dijker@maastrichtuniversity.nl

insufficiently addressed by both the mathematical/game-theoretical and empirical approach to cooperation. Furthermore, it will be proposed that a fruitful evolutionary explanation of cooperation should start with a general consideration of how cooperation is made possible and constrained by general physical and neural properties of behavioral mechanisms. First consider the diversity of behaviors and underlying mechanisms involved in cooperation.

Cooperation may take place largely on the basis of mechanisms involved in self-preservation such as those that motivate and control hunting, feeding, and defense. In many of these cases, individuals are mutually dependent in that they can only satisfy their needs by exchanging goods or services (e.g., flowers trading their attractive nectar with insects or birds for opportunities to distribute their pollen) or coordinating their behavior to realize a common goal (e.g., individuals hunting down a prey animal too large to catch and subdue on their own). Although to an observer these individuals sometimes may appear to behave in "nice", "benevolent", or even "altruistic" ways to each other, it is likely that the underlying mechanisms are not "specifically designed" by evolution to provide fitness benefits to others. Complementarily, given the interdependencies between individuals, there also seems little reason to expect that individuals try to refrain from cooperating or that some form of enforcement would be necessary to prevent them from "cheating" (Clutton-Brock 2009; Leimar and Hammerstein 2010; Maynard Smith and Szathmary 1995; Worden and Levin 2007).

However, some behaviors appear exclusively aimed at improving the fitness of other individuals than the actor, and seem to be made possible by mechanisms specifically designed for this purpose. This can be most clearly seen in cases where recipients of benefits cannot cooperate in the above sense because they are vulnerable or needy, and largely dependent on the care and help of others. Here, benefiting others can only be effective when benefactors are disposed to respond relatively unconditionally and fast to perceived need states (before the other individual's condition deteriorates) and persist in helping until the recipient's situation has improved. Clear examples of these dispositions are the mechanisms underlying parental care which are responsible for behaviors such as feeding, protecting, nurturing, and teaching offspring. It has been argued that similar mechanisms may be involved in helping nonkin in need (e.g., de Waal 1996; Hrdy 2009) or in causing the emotions that would motivate such behavior such as empathy or sympathy (e.g., Batson 1987; de Waal 2008; Dijker 2001).

The mechanisms underlying cooperation are increasingly difficult to distinguish in situations in which mechanisms designed for self-preservation and other-benefiting act together in complex ways. For example, in symmetrical situations in which selfishness or cheating are especially likely, and which game theorists tend to model in terms of the prisoner's dilemma or similar games (see below), individuals may have to be endowed with mechanisms that dispose them to be nice, trusting, forgiving, and tolerant for stable patterns of cooperation to develop (Axelrod and Hamilton 1981; Caporael et al. 1989; Gintis et al. 2003; Richerson and Boyd 2005). Furthermore, although unconditional or speedy and persistent helping may be the most effective in providing benefits to vulnerable or needy others, this behavior may have to be tempered by distrustful or punishing tendencies in order to establish stable patterns of reciprocity (Trivers 1971). The latter tendencies may even be necessary to prevent parents from being parasitized by their offspring (Trivers 1974) or to increase the effectiveness of certain parental behaviors (e.g., aggressive defense of offspring).

The second problem faced by researchers trying to uncover the mechanisms underlying cooperation is more fundamental and relates to assumptions made about the nature of behavioral or psychological mechanisms and the criteria used to recognize and differentiate them. What are the general physical properties that psychological mechanisms must have to enable organisms to effectively influence their environment and to be selected as adaptations? Certainly, these mechanisms should promote inclusive fitness in the long run. But not all conceivable behavioral adaptations seem equally possible, while certain physical properties of adaptive mechanisms seem mandatory. As the above examples suggest, adaptive psychological mechanisms rely on accurate perception or internal representation of certain fitness-relevant properties of other individuals (e.g., their capacity to provide food or safety, their vulnerability or need for protection), motivational properties associated with the reactivity, intensity, persistence, and improvement of behavior, and a particular organization allowing different mechanisms to compete or act together. In addition to effectiveness, they must also allow for selectivity in providing benefits (e.g., on the basis of learned attitudes toward more or less "deserving" recipients of benefits), as is required by genetic cost-benefit models such as kin selection (Hamilton 1964) or reciprocal altruism (Trivers 1971). We have to assume that all these properties of relevant psychological mechanisms heavily rely on using neurons—evolution's most successful behavioral invention—as building blocks.

Unfortunately, the enormously influential mathematical and game-theoretical approach to cooperation (for reviews and general discussions, see Gardner and Foster 2008; Lehmann and Keller 2006; Nowak 2006; West et al. 2007) makes it easy to forget that psychological mechanisms are physical adaptations to a physical environment; whether mental events mediate their behavioral output or not. This

approach does not describe behavioral or psychological mechanisms but "choice strategies" of nonphysical players confronted with a payoff matrix containing values representing fitness costs and benefits. Its goal is to demonstrate (with mathematical proof or computer simulation) that games have particular solutions, termed *Nash equilibria* or *evolutionarily stable strategies*; cells of the payoff matrix representing choices that are best for each player and to which the choices of rational players will converge (for a general treatment, see Binmore 2007a, b). In the next section it is illustrated with a game such as the prisoner's dilemma how difficult it is to establish solely in terms of a game-theoretical approach to what behaviors, let alone underlying mechanisms, choosing rows or columns of a payoff matrix refers. This ambiguity may be partially responsible for the continuing controversies surrounding the definition (West et al. 2007) and evolution of cooperation (Okasha 2010).

The concept of mechanism is also unclearly used in the more empirical approach to cooperation which tries to identify behavioral or psychological mechanisms involved in cooperation on the basis of systematic observations and experiments on real animals (for recent reviews and discussions, see Brosnan et al. 2010; de Waal and Suchak 2010). This approach concentrates on the apparent components of underlying mechanisms; yet, without sufficiently indicating how different components are combined into a single mechanism and different mechanisms fit into a single organism. The long list of mechanistic aspects now thought to be involved in cooperation include references to, for example, cheater detection, empathy, learning, memory, conscience, hormones such as oxytocin, mirror neurons, brain areas (e.g., prefrontal cortex, amygdala, and pleasure or reward center), and a wide variety of mental states (e.g., guilt, sympathy, moral anger) reported by human research participants themselves. In parallel to these references, cognitive psychologists tend to describe psychological mechanisms as sets of instructions or computer programs prescribing how to combine symbols in order to arrive at new symbols or "conclusions"; a process commonly referred to as *information processing* or *computation* (for an influential application to evolutionary psychology, see Cosmides and Tooby 2005; Tooby and Cosmides 1992). However, a description of a psychological process primarily in terms of symbol manipulation is still far remote from a description of true mechanisms with causal and lawful properties (Bechtel and Abrahamsen 2005; Bunge and Ardilla 1987; Sterelny 1990).

Bshary and Bergmüller (2008) recently made the important observation that in the game-theoretical and psychological explanation of cooperation, the term *mechanism* has a different meaning. However, they are not clear about the meaning of the term *psychological mechanism*. For example, they distinguish between "physiological" (oxytocin, brain stimulation) and "psychological" mechanisms (e.g., guilt, pleasure) and contrast these with particular brain structures that would provide the "conditions" for the former two. Yet, from a psychological perspective this is unsatisfactory as a coherent description of a psychological mechanism must be able to integrate neural, physiological, and mental aspects (Bunge and Ardilla 1987). Furthermore, it is also unclear how one should translate choice strategies, the "mechanisms" of game theory, into psychological mechanisms and vice versa.

It should finally be noted that several cooperation researchers working in the game-theoretical tradition have expressed concerns about "cognitive constraints" on the capability of individuals to engage in cooperation or reciprocity (Bshary and Bergmüller 2008; Hagen and Hammerstein 2006; Stevens et al. 2005). However, what they usually mean is that *game-playing* individuals sometimes can only reach particular equilibria when they have sufficient capacity to memorize the previous row or column choices made by themselves and other players, and hence can engage in "book keeping". To what extent these constraints are important in social relationships that are not easily modeled in terms of game theory, is unclear (cf. Silk 2003).

The goal of this paper is to show that a description of the basic properties of adaptive psychological mechanisms and a classification of mechanisms involved in cooperation may considerably improve our understanding of how cooperation and its evolution are constrained by what is physically or neurally possible. Hopefully, adopting a common language for describing behavioral or psychological mechanisms also contributes to bridging the gap between the mathematical/game-theoretical and empirical approach to cooperation.

This paper will not provide detailed descriptions of cooperation in different species or attempt to reconstruct the different routes taken by evolution. Instead, it will focus on general characteristics and broad classes of cooperation and other-benefiting. However, it will especially pay attention to cooperation in humans, for two different reasons. First, many theorists believe that humans are exceptionally good at cooperating, and assume that this is made possible by uniquely human psychological adaptations, associated with properties such as morality, conscience, and tolerance. Yet, an integrative mechanistic description of these qualities is still wanting.

A second reason for paying special attention to human cooperation is that a better understanding of uniquely human psychological adaptations will allow us to better differentiate them from the "more primitive" mechanisms that may be responsible for cooperation in many nonhuman animals. Theories of how the human mind works are often

assumed to be also applicable to explaining social behavior in nonhuman animals. However, an understanding of what the human mind uniquely contributes to cooperation may reveal that "more primitive" psychological mechanisms may be sufficient to cause cooperation in nonhuman animals. At the same time, this awareness may help to appreciate that human cooperation still depends on similar elementary mechanisms and may only *appear* to be determined by uniquely human mental states and processes.

The rest of this paper is organized as follows. After the next section has explained how important it is to distinguish between asymmetrical and symmetrical forms of other-benefiting (i.e., between helping and cooperation), a subsequent section will describe the basic properties of adaptive behavioral and psychological mechanisms. This mechanistic framework will be used to describe in two subsequent sections how mechanisms associated with self-preservation and other-benefiting are involved in helping and cooperation, respectively. The section on helping proposes that a specifically designed psychological mechanism for effective helping (a) is likely to evolve in the context of parent-offspring relationships; (b) acts together with self-preservational mechanisms and acquired attitudes to guarantee selectivity in helping; (c) contributes to the explanation of sympathy, an elusive motivational state generally considered to be the most important mediating cause of helping in humans, as well as other moral emotions such as moral anger, and guilt; and (d) is differentially developed or activated in different individuals and cultures, accounting for important individual and cultural differences in sociality and morality. In the section on cooperation, the mechanisms involved in other-benefiting will be classified, and it will be illustrated how different psychological mechanisms for self-preservation may be combined to improve cooperation. It will also be shown how a mechanism designed for relatively unconditional helping may improve cooperation in highly autonomous agents such as humans, accounting for properties that can be described as morality, conscience, and tolerance. It is finally examined how human evolution has resulted in a parental care mechanism that, supplied with the right perceptual input, can be activated in all individuals, in females as well as males, and in children as well as adults; thereby providing the foundation for a care-based morality.

The Distinction Between Helping and Cooperation

A precondition for distinguishing the different psychological mechanisms that produce behaviors involved in benefiting others is a clear description of the behaviors that are presumably brought forward by these mechanisms. As will become evident in the rest of this paper, one crucial but generally ignored distinction is that between behaviors involved in helping and cooperation. In the social sciences, helping and cooperation tend to be studied in relatively independent research areas, and for good reasons. Consistent with the everyday meaning of the term, social psychologists assume that helping typically takes place in an asymmetrical relationship between individuals in which helping behavior is causally determined by the perception of another individual's vulnerability or need, and the behavior is produced in a timely manner and stops after the recipient's need state has been reduced. This does not deny the influence of a host of other factors on helping such as characteristics of the recipient, the presence of bystanders, or utilitarian motives (for reviews, see Batson 1998; de Waal 2008; Penner et al. 2005).

In contrast, and also consistent with its common meaning (see Tuomela 2000), cooperation refers to a symmetrical relationship in which individuals mutually provide benefits to each other or work together in realizing a common goal (for reviews, see Fehr and Gächter 2000; Hammerstein 2003b; Kollock 1998). Importantly, cooperation is normally assumed to take place among not particularly needy individuals; at least, not that needy that providing benefits primarily would be determined by the perception of another's need (in which case it would be helping). Indeed, a needy recipient of help may not be capable of cooperating in the above sense. As will become evident in a later section, matters are more complex because cooperation sometimes refers to reciprocal helping.

Theorists interested in the evolutionary explanation of other-benefiting behavior often use the terms *cooperation* and *helping* interchangeably. This is unfortunate because one then runs the risk of missing an important physical and neural constraint on the evolution of the mechanism responsible for helping behavior. In particular, one has to assume that in order to help others effectively, a recipient must be capable of causing that behavior in the benefactor; by "unconditionally" activating a reactive sensorimotor or motivational mechanism that produces motor output in a timely, relevant, and persistent manner. It will later be argued that a mechanism for effective helping is most likely to evolve in parent-offspring relationships.

In contrast, cooperation is made possible by a wide variety of mechanisms involved in self-preservation (not to be confused with "selfishness"), although some appear to be "nice" or "specifically designed" to benefit or help others. Importantly, a mechanism specifically designed for asymmetrical helping may come to play a less obvious role in cooperation, but this is difficult to see if that mechanism is not first clearly described.

Especially in the context of game theory, cooperation and helping are difficult to distinguish in terms of their

physical correlates, and this may explain the reluctance of theorists to clearly distinguish them. For example, in the prisoner's dilemma, "cooperation" refers to choosing a column or row of a payoff matrix which, independently of what the other player does, is associated with obtaining the least benefits for oneself, yet a reasonable benefit if the other player would make the same choice (Kollock 1998). One class of explanations explains "cooperation" in terms of the unattractiveness of choosing the "defect" option; for example, by assuming the presence of psychological mechanisms that generate fear of punishment and allow for accurate detection of defectors or cheaters. Another class of explanations centers around the attractiveness of choosing the row or column labeled "cooperate". Although this choice may be explained in terms of self-interested insight ("in the long run, we both will be better off if we choose to cooperate") or other self-interested motives ("to cooperate will help to establish a good reputation in the eyes of the other player or bystanders"), it has also been proposed that it may be caused by a mechanism "specifically designed for" benefiting or helping others in an asymmetrical way; a mechanism referred to with terms such as *niceness*, *benevolence*, *other-regarding sentiments*, or *altruism* (Axelrod and Hamilton 1981; Caporael et al. 1989; Gintis et al. 2003; Richerson and Boyd 2005).

Both classes of explanations have also been combined into descriptions of more complex play strategies that are adopted when the prisoner's dilemma is played repeatedly, of which the most famous is *tit-for-tat*. This strategy assumes that players have a "nice" disposition that causes them to start with a cooperative move, and to subsequently copy the choice made by the other player, thus allowing for both punishment in case the other defects and forgiveness after the other changes from defection to cooperation (Axelrod and Hamilton 1981).

The problem to be noted here is that it is difficult to determine the validity of these and other explanations as long as cooperation is not clearly defined in terms of its behavioral and causal structure. As mentioned earlier, there is nothing particularly benevolent or nice about many forms of cooperation. Furthermore, in the context of the prisoner's dilemma, it is also difficult to see how a player can induce helping by exposing a "nice" player physically to a strong need state or by begging. Similarly, it is difficult to relate game behavior to psychological mechanisms associated with power. That is, one must assume that the use of power to enforce cooperation can only work when benefactor and recipient have complementary sensorimotor or psychological mechanisms that can be mutually activated during the interaction; in such a way that a powerful individual can induce, by signaling physical strength or the potential to inflict harm, a submissive and perhaps fearful individual to provide a benefit. Indeed, such direct causal influences would violate the very principles of game theory which require, among other things, that players make their choices independently and strictly on the basis of information about previous or expected choices of the other player, given the particular payoff matrix (Binmore 2007a; von Neumann and Morgenstern 1944). Of course, a payoff matrix may be constructed to *represent* a particular causal dependency among players and potentials for causal influence. Yet, the way players influence each other is not determined by physical properties that can be perceived while making choices. When economists and psychologists require humans to actually play the games developed by this approach, they similarly allow players to exert as little direct physical influence on each other as possible (Fehr and Gächter 2000; Hammerstein 2003b; Kollock 1998; exceptions are discussed later).

To avoid confusion, this paper uses the general term *other-benefiting* to refer to all behaviors that cause an increase in fitness of other individuals than the actor; irrespective of the particular fitness costs to the benefactor and hence the altruistic nature of the act; and irrespective of the "design" and goals of the underlying mechanisms. At the same time, it is recognized that different kinds of other-benefiting such as helping needy others or cooperation in the narrow sense of working together may be associated with different behavioral mechanisms which may have a different evolutionary origin. Before explaining this, it is important to be clear about the basic concepts that are used to describe behavioral or psychological mechanisms and their functioning.

Physical Mechanisms for Behavioral Adaptation to Physical Properties

An important point about evolution that tends to be obscured by a strong focus on genetic cost-benefit modeling or isolated aspects of underlying mechanisms, or by describing mechanisms primarily in terms of "information processing", is that natural selection works on the physical or material properties of living objects, whether we are dealing with global bodily features, particular organs, or neural mechanisms generating and controlling behaviors that influence the environment. Only when physical properties of bodies and mechanisms allow organisms to effectively adapt to certain physical properties of the world, will they help to promote reproductive success and phylogenetic adaptation. An emphasis on physical properties of individuals and their environment reminds us that, in addition to "designing" adaptations, evolution must also "construct" adaptations on the basis of a limited number of physical possibilities to do so.

Evolutionary Background

The general problem of the genes and evolution in "designing" and "constructing" individual bodies with adaptive properties can be analyzed in terms of three subproblems of adaptation: Self-preservation, reproduction, and benefiting certain other individuals than the actor (cf. Dawkins 1976/1989). Self-preservation refers to the problem that genes and the organisms they code for should have some stability or longevity. At the very least, they should live long enough to replicate successfully. This implies, among other things, that organisms should have adaptive physical and behavioral properties that help them to find and consume food (prey), and to prevent contact with other organisms that have the same goal (predators). Reproduction in its sexual form requires finding a suitable mate. At a minimum, this mate should have "good" or "healthy" genes that help to code for properties that increase the chances of the new combinations of genes, and the offspring they code for, to self-preserve and reproduce. Finally, other-benefiting refers to the problem of *not* letting self-preservation goals prevail during contact with organisms carrying copies of one's genes, such as the products of previous reproduction (offspring) and other kin. That is, kin should not be eaten and not be damaged in the process of defending oneself against *their* self-preservation needs. In addition, other-benefiting should result in properties that help promote the self-preservation and reproductive success of offspring and other kin. At least in relation to offspring, effective other-benefiting behavior may directly promote parents' reproductive success and therefore requires mechanisms that are "specifically designed and constructed" to generate that behavior.

The three classes of adaptive behavioral mechanisms each respond to a particular class of properties of objects. Mechanisms associated with self-preservation have to respond to properties such as the dangerousness and controllability of predators, or edibility of food. Mechanisms relevant for reproduction should respond to properties associated with sexual receptivity and potential contribution to the fitness of offspring. Of special interest in this paper is the main property of objects to which a behavioral mechanism devoted to other-benefiting should respond: Vulnerability. In the case of living things, vulnerability refers to the likelihood or disposition of individuals to change into a state of decreased fitness, inconsistent with their "design specifications", when exposed to certain conditions (e.g., strong outward forces causing deformation or disintegration). Examples of cues associated with vulnerability are relatively small size, transparency or light color indicating relatively small mass or brittleness and, when in motion, relatively less potential force would the object impact on another object with more mass and/or counter force. Opposite to vulnerability is physical strength and potential force, causing relatively vulnerable objects to deform or disintegrate upon collision or handling. Individuals concerned with preventing this from happening should be expected to relate vulnerability and physical strength in complementary ways to each other. For example, behavior that takes into account an object's vulnerability may consist of protecting the object against strong outward forces, and applying just enough force to gentle handle or manipulate it in goal-relevant ways (Dijker 2008; Gorniak et al. 2009).

In general, effectively responding to fitness-relevant properties requires that behavior is determined by the accurate perception of these properties. A basic problem, however, is that an object's properties cannot be directly perceived on the basis of sensory input alone. After all, a property is a disposition of an object to change from one state into another when the object is exposed to certain conditions. Hence, a property can only be discovered through influencing the object and observing the consequences, which may result in a representation of, or memory for, the property in terms of an "if–then" relationship. For example, an object's vulnerability can be internally represented by the expectancy that the object will be deformed or break when manipulated in certain ways, and its heaviness by the expectancy that a particular intensity of motor output or force is needed to overcome gravitational force.

In early organisms, accurate perception and adaptive behavior were one and the same thing. That is, these organisms acquired simple reactive sensorimotor mechanisms of which the sensors started to activate effectors at the moment they were activated by particular input stimuli that were reliably correlated with fitness-relevant properties. The nerve cell or neuron connecting a simple sensor to an effector has proven evolution's most successful tool to build these adaptive mechanisms.

Consider, for example, a neuron that connects light-sensitive cells to effector cells (e.g., the muscle cells of a flipper) so that behavioral output is directly caused and directed when the sensors receive relevant input. When the light-sensitive cells are stimulated by a light source and light is correlated with the presence of food or the right temperature, a simple sensorimotor mechanism may evolve that causes the organism to swim toward the light source (i.e., "motivates" the organism to produce motor activity until a certain endstate or goal has been realized). The more reliable the correlation, the more the neural mechanism or disposition contributes to the organism's fitness and to preserving its genetic code for future generations.

By increasing the number of neurons involved in linking sensors to effectors, and by varying the connection strengths between them, the organism may acquire the

capacity to engage in one of the most elementary forms of learning, thereby improving its performance during its lifetime. In particular, through classical conditioning, new or conditioned stimuli, originally not able to do so, may gain the causal potential to activate the mechanism, thus increasingly allowing the organism to respond in an anticipatory fashion (e.g., arriving at the food location before food becomes available).

Behavioral adaptation will be further improved by a neural network structure in which sensorimotor mechanisms performing different kinds of functions are allowed to function relatively competitively or independently from one another through mutual inhibition, thus resulting in a process similar to decision making (Grossberg 1980; Ludlow 1980). For example, it would be adaptive that a system for seeking food near a light source competes with a system specifically responding to stimuli associated with the presence of predators. In this way, strong stimulation of the latter system can inhibit food seeking behavior and induce the organism to change to an escape mode in order to avoid being eaten.

However, a greater improvement in the organism's capacity to produce adaptive behavior will occur when the neurons mediating sensors and effectors can form increasingly accurate and stable internal representations of objects and their invariant properties. The general way in which neural networks or brains can do this is by changing the connection strengths between neurons on the basis of sensory feedback about the consequences of the organism's own behavior; a learning process (operant conditioning) resulting in the formation of sensorimotor networks that represent correlations between sensory and motor activity or what-leads-to-what expectancies. Many theorists assume that it is through these kinds of internal representations that the brain is capable of perceiving invariant or constant properties of the environment (Gregory 2005; Grossberg 1980; Helmholtz 1878; Jeannerod 1994; O'Regan and Noë 2001; Shepard 1989; Sommerhoff 1974). For instance, whereas light rays in the above example are cues that can unconditionally trigger the behavioral mechanism, and the reactive mechanism "motivates" behavior, interaction with, or manipulation of, objects present at the location of the light source may reveal to what extent these objects are edible when manipulated in a particular manner.

A complementary way to view the combined activation of sensorimotor or motivational systems and internal representations of properties that mediate between sensors and effectors is to see them as motivational states or emotions that continue to drive and control behavioral output when sensory stimulation is interrupted (Bindra 1985; Grossberg 1980; Lewis 2005; Toates 1986). Because these states allow organisms to fine-tune their behavior to objective properties of the world, they will result in increasingly effective action. A similar view of the production of adaptive behavior underlies "embodied" approaches to human cognition (Clark 1999). For example, it is recognized that in building robots it is necessary to start with simple reactive mechanisms that, through learning, construct increasingly elaborate and accurate internal representations of their environment without the need of instructions and explicit symbolic representations (e.g., Brooks 1997; Bryson 2000; Sloman and Chrisley 2005).

Once stable internal representation of properties is possible, it starts to make sense to use symbols to refer to properties in the outside world in a "truth-conserving" manner, as well as to refer to relationships among properties and to logically derive new ones. Probably a typically human property of brains is that they can use an advanced symbol system (e.g., language) to describe and think *about* the internally represented properties of the world (and of the body responding to this world), and derive new properties on the basis of reasoning and logic. This enables humans more than any other species to consider different courses of action without or before experiencing their (sometimes fitness-reducing) consequences.

Unfortunately, it is this symbolic capacity of the human mind that also forms the major obstacle to understanding how brains work and how they evolve. In particular, our own ability to describe events and properties in terms of symbols and rules for combining symbols (e.g., language), and to simulate processes of symbol manipulation on digital computers, have led psychologists to equate the concept of psychological mechanism with a set of instructions or computer program. This may obscure the tremendous capacity of the brain to internally represent the world and its properties in a quite accurate and stable manner on the basis of highly variable sensory input, and without the use of symbols.

A Model for Adaptive Psychological Mechanisms

In a well-adapted organism, all three ways of responding to properties of the world and its objects are integrated. This is illustrated in Fig. 1. At first, an organism responds to the properties of objects on the basis of cues or trigger stimuli that are correlated with these properties. These cues trigger sensorimotor or motivational systems (which compete for expression) that generate goal-directed behavior, resulting in changes in objects and the relationship between organism and objects. On the basis of sensory feedback from these changes, stable internal representations of properties and objects are formed that, together with activated motivational systems, determine the motivational state of the organism. Attitudes are internal representations of objects that are associated with unconditioned cues that have the potential to trigger motivational systems. Hence, activation

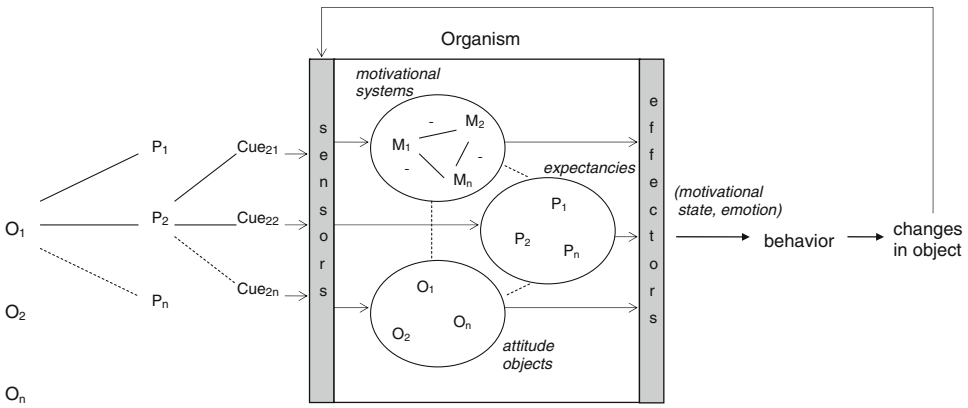

Fig. 1 Production of behavioral responses to the properties (*P*) of objects (*O*) on the basis of cues and feedback from self-produced changes in objects. Properties are internally represented in terms of expectancies. Motivational systems compete with each other through inhibitory connections

of these representations may also activate motivational systems and exerts a biasing influence on motivational systems involved in other-benefiting and self-preservation. In Fig. 1, negative attitude objects could be represented by an excitatory connection between a represented object and the motivational systems responsible for fear or aggression, together with connections to negatively evaluated properties or negative expectancies. Positive attitudes could be represented by an excitatory connection between a represented object and a motivational system underlying (parental) care, together with connections to positively evaluated properties (for a similar psychological description of attitudes, see (Cacioppo and Berntson 1994; Staats 1968). (Of course, there may be other motivational systems that may be linked to positive attitude objects and positive expectancies such as the ones involved in feeding and sex.) As argued below, the biasing influence of attitudes on the activation of motivational systems allows them to play an important role in selectivity in other-benefiting behavior.

The activation thresholds of motivational systems may also be lowered by means of priming; previous or contextual activation of these systems (e.g., by cues unrelated to the current object or situation) may bias them to respond earlier to new stimulation.

Figure 1 also suggests that the same property may be associated with different cues. For example, the dangerousness of an object (P2) may be associated with its loomingness, particular shape, or alarm calls of other objects that are present. The vulnerability of an object may be associated with age-related immaturity cues such as small size and particular behavior. A situation may activate more than one motivational system. For example, a vulnerable but happy child may activate a motivational system dedicated to care, yet when it shows expressions of distress and alarm it may also activate fear or aggression. Figure 1 shows that motivational systems, representations of properties, and attitudes are closely connected and that all contribute to generating motivational states or emotions. Finally, to incorporate a role for symbolic representation and symbol manipulation in Fig. 1, imagine that the input, output, and functioning of the mechanisms can be described in terms of a symbol system (e.g., language) so that the organism can "think about" them, and that these thoughts sometimes also have the potential to activate motivational systems and hence cause behavior.

Implications for Explaining Helping Behavior

The literature on helping and cooperation shows a strong tendency to analyze other-benefiting in terms of selectivity; selecting the "right" and rejecting the "wrong" or less deserving recipient of help, or differentially investing in the two. This aspect has been most strongly emphasized in two influential genetic cost-benefit models of social behavior. In particular, Hamilton's (1964) theory of kin-selected altruism or inclusive fitness proposes that, at the level of genes, fitness can be promoted by benefiting others if benefactors select recipients that are sufficiently genetically related. Trivers (1971) argued that the evolution of a trait for benefiting genetically unrelated individuals is also possible as long as the benefactor selects recipients that are likely to provide return benefits (i.e., has a good reputation as a reciprocator) and hence the costs of helping can be compensated during the benefactor's lifetime. Hence, both models predict the evolution of psychological mechanisms for selecting the right recipients of benefits.

Unfortunately, this emphasis on selectivity has obscured the importance of the effectiveness of helping. Using the present mechanistic approach, it can be shown that effectiveness depends, among other things, on behavioral qualities such as speed, persistence, and "unconditional kindness"; and that these properties are made possible by the automatic or unconditioned activation of a motivational mechanism by cues that are correlated with the vulnerability and neediness of a potential recipient of benefits. In this section, it is first argued that a mechanism for effective helping having these properties is especially likely to

evolve in the context of parent-offspring relationships where it contributes unambiguously to reproductive success. It is subsequently argued that unconditional helping does not exclude selectivity and that both are made possible by letting motivational systems for unconditional helping and self-preservation act together. In another subsection, it is shown how a better understanding of the mechanism underlying effective parental care helps to clarify the nature of the elusive motivational state thought to be centrally involved in motivating helping behavior in humans (and perhaps in several nonhuman primates as well): Sympathy, as well the nature of many other moral emotions such as moral anger and guilt. Finally, it is argued that individual and cultural differences in sociality depend on the likelihood with which mechanisms for parental care and self-preservation can be activated in different individuals and cultures, warning us that our explanations of helping and morality in general are dependent on the individuals and cultures selected for study.

Parent-Offspring Relationships as the Context for the Evolution of a Mechanism for Effective Helping

Kin-selection or inclusive fitness theory (Hamilton 1964) not only predicts that a mechanism for effective other-benefiting is especially likely to evolve in a relationship between parents and offspring, but also in response to a specific physical property of the recipient of help. In particular, the theory predicts that such a psychological mechanism can evolve when degree of genetic relatedness between benefactor and recipient is high and the ratio of costs to benefactor and benefits to recipient is relatively small. These requirements tend to be satisfied for parents providing benefits that are both extremely beneficial to vulnerable offspring and of little costs to the benefactor. The key determinant of such a small cost-benefit ratio would be a capacity of parents to accurately perceive and appropriately respond to the property of vulnerability; the likelihood or disposition of individuals to change into a state of decreased fitness when exposed to certain conditions.

There are different ways in which extremely vulnerable offspring are especially likely to profit from parental behavior with a small cost-benefit ratio. For example, they may profit from their parents doing little more than simply staying around and *not* harming them, thereby also discouraging others who might be interested to do so (e.g., predators). Vulnerable offspring may also profit from a host of low-cost side-effects of parents' concern with their own self-preservation. For example, in gathering food for themselves or in self-protection, parents may simultaneously supply food or protection to offspring. These examples suggest that, if an effective mechanism for other-benefiting is to evolve, it may sometimes have to work together with mechanisms involved in self-preservation (e.g., in aggressively defending offspring) and sometimes to compete with them (e.g., self-protection may be more urgent than helping others).

However, in the context of a relatively stable attachment between parents and offspring, and via a process of mutual adaptation or evolutionary arms races (Dawkins and Krebs 1979), parental mechanisms may evolve that are increasingly responsive to specific changes in the fitness or well-being of offspring. For instance, if some parents are inclined to do a little more than staying around (e.g., more actively scare away predators, bring more food back than can be consumed alone and leave it for offspring to consume), they would have an advantage in terms of inclusive fitness. Complementarily, if some newborns would manage to trigger these new mechanisms more effectively than others (e.g., by wandering away and alerting predators, by begging for food), their inclusive fitness would also increase. In turn, parents will make their care more effective (e.g., prevent the offspring from wandering from the nest, regular provision food), while offspring become more proficient in soliciting care (Kilner and Johnstone 1997).

As outlined in an earlier section, the brain can adaptively and effectively respond to fitness-relevant properties such as the vulnerability of offspring by responding in a reactive and relatively unconditional manner to sensory input correlated with vulnerability. Second, "motivated" by these reactive behavioral mechanisms, more advanced brains can form increasingly accurate internal representations of an object's vulnerability. Finally, a capacity for symbolic representation and symbol manipulation ensures that parental care will be increasingly associated with particularly low fitness costs and high effectiveness or benefits to recipients. In addition to recognizing the mere physical features of immaturity, this capacity allows for vulnerability of offspring to be represented in multiple symbolic ways. Vulnerability may, for example, be implied by a vulnerable posture (e.g., lying on the ground), a particular situation (e.g., a single individual confronted with an angry crowd), or a social relationship in which individuals are dependent on trust and mutual willingness to cooperate given strong temptation and opportunity to harm each other or cheat (see below). Complementarily, this capacity enables individuals to imagine different behaviors that would be effective in preventing harm or providing fitness opportunities. Improved intelligence also enables individuals to provide care in a critical and punitive manner, preventing them from being victimized by social parasites while continuing to provide care in a timely and relatively unconditional manner to those that urgently need it.

More specifically, mammalian parents probably were the first to experience many of the so-called *moral*

emotions that are currently being distinguished by psychologists such as sympathy, guilt, or moral anger. As explained below, these emotions can be conceived as adaptive motivational states, tailored to specific changes in fitness of vulnerable offspring and their attributed causes, allowing parents to fine-tune their care to widely different situations in terms of, for example, aggressive defense, cleaning, healing, punishment, education, or simply tender physical contact.

Individuals are also Selective in Choosing Recipients of Benefits

Effectiveness of helping does not exclude selectivity. The contrary, the evolution of an effective mechanism for parental care is based on extreme selectivity, ensuring that parents primarily invest in the products of their own reproductive efforts. As described in relation to Fig. 1, selectivity is best described in terms of attitude formation and functioning. That is, a strong activation of a parental care system may result in a strong positive attitude and attachment, which will bias the activation of the system on subsequent occasions. Activation of self-preservational mechanisms (e.g., because of nonreciprocation or other harmful features of the recipient or situation) may result in negative attitudes, exerting an inhibitory influence on other-benefiting. This view is consistent with the well-established facts in social psychology that we are especially likely to help positive attitude objects such as kin, friends, or reciprocators (Korchmaros and Kenny 2006; Penner et al. 2005) and refuse to help or punish negative attitude objects such as disliked others or social parasites (Miller et al. 2003).

It is important to note, however, that a too strong emphasis on the role of selectivity and attitudes may have obscured an important question about the nature of other-benefiting behavior: What properties should a behavioral mechanism have to generate relevant and effective other-benefiting behavior that really makes a difference for the other individual, whether one feels attached to this individual or not?

On the Right Interpretation of Sympathy and Other Moral Emotions

It has been argued (Dijker 2010) that the different motivational states or emotions that mammalian (and especially human) parents experience in relation to immature young, can be conceived of as so-called *moral emotions* that all depend on the activation of a parental care system by a vulnerable object (which then changes into an object of care), together with activation of systems involved in aggression and fear. In particular, these emotions are adaptive motivational states, tailored to specific changes in fitness of vulnerable offspring and their attributed causes, allowing parents to fine-tune their care to widely different situations. For example, moral anger is caused by the perception that a third party threatens the fitness of the care object and results in aggressive defense and punishment; sympathy is caused by the perception that the fitness of the care object has actually decreased but that a responsible agent is absent or irrelevant, resulting in comforting and healing of the object. Dependent on the attributed cause of the change in fitness, aggression may also be directed at the object itself (to induce better self-care) or the self, thus resulting in guilt (Nelissen and Zeelenberg 2009). Once again this illustrates the point that, in order to produce effective outcomes, a mechanism specifically designed for other-benefiting has to act together with mechanisms involved in self-preservation. Below, it will be argued that this advanced system of emotions may be exapted for reciprocity and cooperation among not particularly vulnerable or needy adult strangers. Although generally acknowledging that many moral emotions are based on care, other theories of moral emotions (for reviews, see Haidt 2003; Tangney et al. 2007) are less clear about the underlying psychological mechanisms.

Here, this theory of moral emotions is used to solve an urgent problem in the psychological explanation of emotions that motivate helping behavior in a rather altruistic manner and which have been variously termed *sympathy*, *pity*, or *empathy*. Darwin (1872/1998, p. 215) found it relatively self-evident that sympathy is commonly felt in response to the suffering of those that we feel attached to (positive attitude objects), but struggled with sympathy felt for total strangers. An influential solution to this problem, proposed by psychologists (e.g., Batson 1987; Preston and de Waal 2002), is to assume that especially humans are capable of taking the perspective of needy individuals and that this would trigger the emotion of sympathy and a tendency to help; even in the absence of attachments or positive attitudes. (This is the reason why empathy often is seen as an emotion and equated with sympathy.) This account fails, however, to explain why, in taking the other's perspective, we *care about* improving the other's situation and spend much energy and time to do so (Wispé 1991). Here, it is proposed that the parental care mechanism not only is the main determinant of sympathy with suffering individuals but also of a general tenderness and softness in response to others who do not (yet) suffer but are vulnerable and fit (e.g., a smiling, active, and curious infant). Sympathy is a motivational state in which tenderness combines with the distress felt when observing an actual decrease in fitness of the vulnerable care object (Dijker 2001, 2010; McDougall 1908/1948), and should not be confused with empathy. It is a state automatically

triggered by the reactive mechanism of parental care and an inability to attribute harm to external causes. Yet, perspective taking may help potential benefactors to better understand the nature of the other's need and the behavioral alternatives for improving the needy individual's situation.

The parental care system may not only be responsible for a wide variety of moral emotions but also for social qualities that have been termed *tolerance*, *patience*, and *gentleness*, and which may also play an important role in symmetrical cooperation (see below). Importantly, these qualities and the care-based moral emotions may become less visible in individuals and groups in which self-preservational mechanisms inhibit the activation of a care mechanism. These individuals and groups are more likely to morally respond on the basis of fear, aggression, submission to dominant others, and shame (for a discussion of differences between shame and guilt, see Tangney et al. 2007).

Individual, Group, and Contextual Differences in Helping

The extent to which motivational systems associated with self-preservation and other-benefiting can influence social behavior depends on the relative ease with which these mechanisms can be activated in individuals, species, and societies. That is, some individuals, species, or societies are more likely to interact on the basis of a strongly activated care system, whereas others regulate social interaction primarily on the basis of a strongly activated system responsible for fear and aggression and likely responsible for punishment, obedience, and public shaming or stigmatization. This implies that one should be careful not to draw conclusions about *the* psychological mechanisms that would be responsible for cooperation on the basis of observations made on a single group or culture.

With respect to individual differences in helping behavior in mammals, one would first of all expect sex differences. In particular, in mammals, females are more likely to have a stronger developed and more easily activated care system, and less developed fight-or-flight system, than men. In humans, men are not only less nurturant (e.g., Costa et al. 2001) but also less likely to adopt an ethic of care than justice (e.g., Jaffee and Hyde 2000; Skoe et al. 2002), and respond with more anger and punishment to offenders of norm violations (e.g., Gault and Sabini 2000) than women.

In humans, individual differences in care and self-preservation also express themselves in a wide variety of personality characteristics relevant for social behavior. For example, people high in authoritarianism or conservatism show a tendency to fearfully submit to powerful others, and to aggress or show contempt to those who are disobedient and violate norms (Altemeyer 1998). In contrast, liberalism or egalitarianism is associated with a morality of care and greater "softness" in responding to deviance, expressing itself, for example, in greater sympathy with needy or disadvantaged others (for a review, see Dijker and Koomen 2007). In making moral judgments, liberals find notions of care and reciprocity more important than of obedience and ingroup loyalty, while increasing conservatism is associated with a decrease in relevance of the former two but an increase in relevance of the latter two moralities (Haidt 2007).

Societies or cultures can be seen as sets of stimuli that exert a chronic (priming) influence on the different mechanisms involved in cooperation. In particular, human societies in which the care system is chronically activated can be associated with an egalitarian, and societies in which the fight-or-flight system is chronically activated, with a hierarchical social organization (often accompanied with a reproductive system based on male dominance). In combination with collectivism and small group size, the former societies are known to be relatively communal, peaceful, forgiving, and lacking in intergroup conflict; the latter (especially in combination with larger group size and greater complexity) as more punitive, stigmatizing, and engaging in intergroup conflict (Boehm 1999; Hofstede 2001; Knauft 1991; for discussions, see Dijker and Koomen 2007).

An egalitarian social structure has returned in modern Western society, yet this time in combination with strong individualism, division of labor, and trade (Inglehart and Baker 2000), and a variety of institutions and strategies to suppress negative behavior toward a wide array of deviant characteristics of interaction partners (Dijker and Koomen 2007).

Finally, we should expect that relatively temporary circumstances also influence the likelihood with which social life and cooperation in particular will be influenced by mechanisms associated with other-benefiting and self-preservation. For example, war, famine, or plagues will strongly activate the mechanisms involved in self-preservation, thereby resulting in relatively distrustful and punitive kinds of social interaction.

Implications for Explaining Cooperation: Towards a Classification in Terms of Psychological Mechanisms

The search for adaptive mechanisms underlying cooperation has been strongly determined by a tendency to first make a distinction between mechanisms involved in obtaining indirect and direct fitness benefits, and then trying to classify the different mechanisms within these broad

categories. Direct benefits refer to benefits obtained by benefactors during their lifetimes, whereas indirect benefits refer to benefits to genetically related recipients; i.e., benefits that increase the inclusive fitness of benefactors but typicaly imply fitness costs that are not compensated during the benefactors' lifetimes (Hamilton 1964; Trivers 1971; West et al. 2007). Research on psychological mechanisms involved in obtaining indirect fitness benefits have usually been restricted to mechanisms for determining the degree of genetic relatedness between benefactors and potential recipients of benefits. One can be sure, however, that theorists interested in kin-directed other-benefiting often have asymmetrical helping behavior in mind, like feeding starving, or defending vulnerable kin; and that they would agree that this behavior requires "specific design" in order to be effective.

In contrast, the behaviors thought to be involved in obtaining direct benefits are much more divers, motivating theorists to construct increasingly fine-grained classifications in terms of, for example, reciprocity, pseudoreciprocity, indirect reciprocity, or by-product mutualism (Bergmüller et al. 2007; Lehmann and Keller 2006; Nowak 2006; Stevens et al. 2005; West et al. 2007). Yet, it is not always clear to what extent these classes refer to unique or common mechanisms, keeping the true sense of psychological mechanism in mind. Most importantly, it is not clear if mechanisms involved in obtaining indirect benefits may also be involved in obtaining direct benefits and vice versa.

In the present paper, the classification of other-benefiting behaviors that will be proposed is unambiguously in terms of true behavioral or psychological mechanisms. Specifically, these behaviors will be attributed to mechanisms that, in principle, can be specified in terms of underlying brain mechanisms for connecting sensory input to motor output, and to discrete physical and fitness-relevant properties of other individuals and the common environment; properties associated with sensory cues that can activate mechanisms. Because this description of a mechanism establishes a lawful connection between perceptual input and motor output, its explanatory use qualifies as a causal explanation (Bechtel and Abrahamsen 2005; Bunge and Ardilla 1987).

The goal of the classification is to differentiate psychological mechanisms, to examine how they may be combined, and, to understand how a psychological mechanism for unconditional helping may improve symmetrical cooperation. No attempt will be made to determine for specific other-benefiting behaviors whether they represent (indirect) reciprocity, pseudoreciprocity, or by-product mutualism. Instead, it is proposed that one should first be clear about the basic psychological mechanisms that animals have at their disposal before one attempts to assign other-benefiting behaviors to these classes. Interpretation problems repeatedly occurring in previous classification attempts (e.g., whether previous examples of reciprocity should be better considered pseudoreciprocity or whether reciprocal interactions are really based on cost-benefit calculations or on attitudes) indicate that an analysis in terms of true mechanisms and their flexible combination is urgently needed.

Classification of Other-Benefiting Behaviors

Figure 2 distinguishes three main classes of other-benefiting behaviors. At the far left, bottom row, two mechanisms are listed that are activated and partially controlled by the recipient of the benefits. Helping is asymmetrical other-benefiting in the sense that it is caused by the recipient's condition and/or the active soliciting of help; on the basis of a sensorimotor and motivational mechanism in the benefactor that establishes a lawful connection between perceptual input and motor output. As suggested above, the

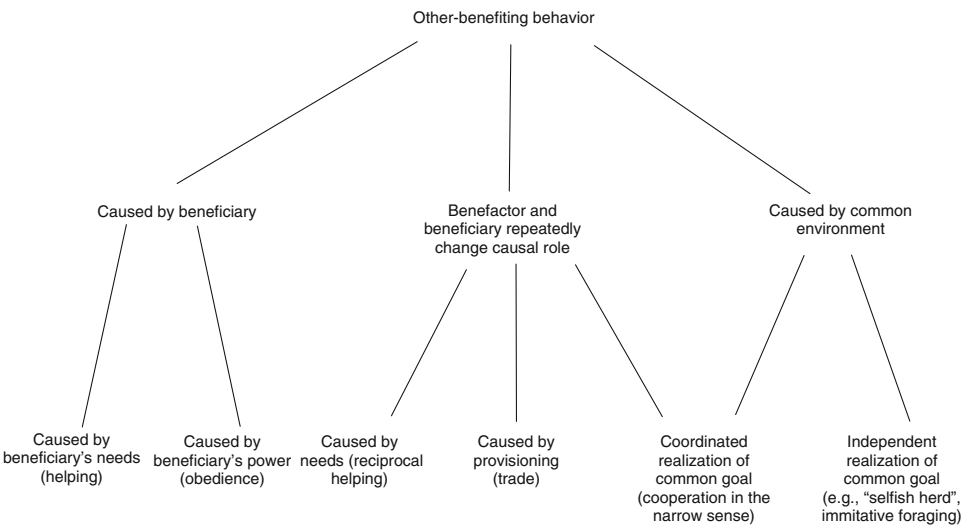

Fig. 2 A classification of other-benefiting behaviors in terms of different causal antecedents and proximate mechanisms

responsible psychological mechanism most likely originates from the evolution of parental care.

As will become clear later, it is useful to distinguish asymmetrical helping in species in which behavior is largely produced by reactive sensorimotor mechanisms, and in species in which individuals are relatively autonomous and behavior is produced with much more mediation by internal representation and mental processing. Ants would be a good example of the former species and it may be speculated that their "cooperation" almost entirely is based on reactively produced helping responses to the perceived vulnerability and needs of others, resulting in a "super organism". In particular, in most members of an ant society, other-benefiting prevails over self-preservation and reproduction as (dependent on their particular role) these members engage in behavior such as caring for the young, self-sacrificial defense, carrying siblings, and offering themselves as living instruments or components for the construction of nests, bridges or floating devices (Hölldobler and Wilson 1990). The fact that the queen is cared for by her offspring does not contradict the present interpretation that most ant cooperation may be derived from kin-directed parental care if it assumed that she triggers care by emitting cues that are correlated with her vulnerability and chronic dependency on her offspring.

While helping behavior in highly autonomous agents such as humans may still depend on a reactive parental care mechanism, its effects are likely to depend much more on competition with self-preservational mechanisms and the enormous mental capacity to trigger these mechanisms and to think about ways to satisfy them. However, the combination of a strong care mechanism and mental capacity in humans may also be responsible for the probably uniquely human property of conscience, enabling the extensive cooperation in autonomous individuals (see later).

The relatively unconditional nature of a mechanism for helping is most clearly suggested by the existence of interspecific parasitism. For example, Trivers (1985, pp. 50–51) describes how beetles of *Atemelis pubicollis* try to get the attention of ants of *Myrmica laevinodis* by tapping them on their mouth parts with their antennae and front legs (behavior which can be interpreted as begging), resulting in the ant regurgating food. A more familiar example would be the tendency of some birds to make use of the parental care mechanism of other birds by laying their eggs in the latter's nests.

Yet, another form of asymmetrical other-benefiting occurs when the recipient has the power to harm or punish the benefactor, thus causing the benefactor to provide benefits in order to avoid punishment (Clutton-Brock and Parker 1995). The main responsible mechanism would be the one involved in escape and aggression. For example, dominant individuals may suggest (e.g., through increasing in size or aggressive displays) to submissive individuals that they have the potential to physically harm them, whereas the latter may suggest (e.g., through making themselves small or delivering the demanded goods or services) that they are harmless and even ready to escape. It is interesting to note that parents may employ asymmetrical helping and power simultaneously to effectively benefit their children (for a discussion of parental styles in primates, see Maestripieri 1999).

A second major class of other-benefiting behaviors shown in Fig. 2 includes symmetrical or reciprocal social behavior, with individuals engaged in a process in which they provide benefits to each other. In one subclass, individuals respond to each others' needs and are essentially engaged in helping or need reduction; in the other, they provide unsolicited fitness opportunities to others and may even create needs in others. The first subclass is discussed by Trivers (1971) in terms of reciprocal altruism. Although often not made explicit, Trivers starts from the assumption that mechanisms for benefiting needy others in relatively unconditional ways are already in place before reciprocity can evolve. This is clear from examples like "helping in times of danger [e.g., accidents]" and "helping the sick, the wounded, or the very young and old" (1971, p. 45). It is also evident from his examples that effective helping often is crucially dependent on the willingness of the donor to immediately and without hesitance engage in quite risky behavior. Finally, Trivers suggests a competition between unconditional helping and aggressive tendencies by noting that anger is not only necessary to punish and discourage cheaters, but also to inhibit strong and indiscriminate helping tendencies that are associated with the experienced pleasure of helping others (1971, p. 49). It must be noted that there is remarkably little evidence for the existence of reciprocity in nonhuman animals (Clutton-Brock 2009; Hammerstein 2003a). It is safe, however, to conclude that in humans, reciprocity is considered a major value given the importance attached to norms prescribing it (Gouldner 1960) and the function of gratitude (McCullough et al. 2001; Trivers 1971).

A parsimonious way to conceive of a psychological mechanism that allows individuals to temper their relatively unconditional helping tendency is to assume that this tendency can compete with an independently functioning aggression system that is also responsible for the formation of conditioned negative attitudes towards social parasites (Stijnen and Dijker in press).

A second subcategory of symmetrical other-benefiting involves the mutual provisioning of benefits not caused by the perception of others' needs but by an expected return benefit. By making available a fitness opportunity to others, the benefactor simultaneously takes the opportunity to consume a fitness benefit made available by the recipient,

situation, or bystanders. For example, flowers may exchange their attractive nectar with insects or birds for opportunities to transport their pollen to other flowers (for other examples, see Clutton-Brock 2009; Connor 1995). Providing benefits to a recipient to obtain return benefits from bystanders, termed *indirect reciprocity* or *social prestige*, falls in the same subcategory.

A tremendously important new possibility for mutualism is born when individuals are capable of intentionally producing surplus goods or services at relatively low costs, to be exchanged with others who intentionally produce different surplus goods or services. Why not catch at relatively little additional costs several more prey animals, make several more arrow heads, or cook dinner for two instead of one, if you can trade these surplus products for goods and services that are currently more useful to you? A famous description of this trading process, primarily guided by self-preservational motives, and assumed to be the basis of human division of labor and economic growth, was offered by Adam Smith (1776/1910). Thus in order to obtain others' goods or services, "[Man] will be more likely to prevail if he can interest their self-love in his favor and shew them that it is for their own advantage to do for him what he requires of them" (Smith 1776/1910, p. 13).

In psychological terms, this process can be described as the mutual rewarding of each others' behavior by means of operant conditioning; with the behavior not specifically directed at improving others' fitness but simply performed to obtain rewards.

The third major class of other-benefiting behavior in Fig. 2 is caused by common environmental properties that activate a self-preservational or reproductive mechanism present in the different individuals involved. Important reasons for individuals to flock, herd, or school is that this allows them to decrease the likelihood of becoming targeted by a predator, increase the likelihood of obtaining food or mates, or safely give birth (Hamilton 1971; Wilson 1980). Note that in explaining herding, individuals are not expected to specifically provide costly goods or services to others; they simply are assumed to want to be in each others' presence to obtain benefits from herding. The common goal of safety or obtaining food is realized by members of the herd independently directing their behavior at the same environmental properties such as the presence of a predator.

Cooperation in the narrow sense of working together is another subclass of other-benefiting caused by common fitness threats or opportunities that activate mechanisms associated with self-preservation such as feeding, escape, or aggressive defense. However, unlike the former subclass, cooperation in the narrow sense is additionally caused by active communication, coordination, and role differentiation (Noë 2006). Typically, individuals engage in this sort of activity because they cannot realize the endstate on their own and hence need the contribution of others to obtain it. This process can also be described as *synergy* (Alvard 2001; Maynard Smith and Szathmary 1995). Hunting down prey animals too big to catch on one's own would be a good example, with different individuals being responsible for different tasks (Packer and Ruttan 1988). Thus cooperation in the narrow sense differs from helping in that the other individual's need state is not a sufficient cause for the coordinated activity to take place. (In fact, a serious need state would make it difficult for individuals to effectively contribute to obtaining a common goal.) This makes it also different from Trivers' reciprocal altruism according to which truly asymmetrical helping takes place repeatedly. (Unfortunately, many later references to Trivers (1971) do not clearly distinguish reciprocal helping from cooperation in the narrow sense and from exchange among not particularly needy individuals.)

Another important aspect of this kind of mutualism is that it often does not make sense for interaction partners to "cheat" or contribute less than the other, as it may be physically impossible to realize the common goal without a specific distribution of individual efforts (Clutton-Brock 2009; Leimar and Hammerstein 2010; Maynard Smith and Szathmary 1995). This may also hold for exchange or trade, when individuals have to use each other reciprocally as instruments to realize self-preservational goals.

The present classification suggests that different psychological mechanisms may be combined to make cooperation in the narrow sense increasingly effective. Consider the following possibilities. The environment may function as a common cause for cooperation, for example, by presenting predators or prey to a group of individuals. When the relevant mechanism for self-preservation is activated in these individuals, they will collectively orient towards the predator or prey, resulting in collective defense or hunting, respectively. While this may initially merely involve a kind of imitative defense or imitative foraging (Wilson 1980), it may increasingly make use of communication between individuals. In particular, in collectively trying to realize the common goal, individuals may (a) communicate temporary need states and solicit assistance to perform the tasks associated with their role well (helping), (b) force others to obey instructions (power), and (c) provision and exchange unsolicited services and goods, thereby preventing future need states and drop-out of cooperators.

Improving Cooperation with Asymmetrical (Parental) Care

For different reasons, cooperation among individuals endowed with brain mechanisms for relatively unconditional

and asymmetrical (parental) care can be considerably improved. First, cooperation may be almost entirely controlled by mutually activating reactive mechanisms derived from parental care, as is perhaps true for eusocial insects such as ants.

Second, when individuals are temporarily incapable of working together with others due to injury or illness, it pays to allow them to get better or to actively attempt to heal them. In all human societies, a caring tendency has developed into an elaborate system of healing and medicine, together with a set of normative prescriptions to prevent social parasitism, known as the *sick role* (Fábrega 1997; Parsons 1951). The underlying motivation is apparently so strong that it also has resulted in a variety of institutions for specialized care for the permanently ill or disabled.

Third, to make symmetrical cooperation and reciprocity on the basis of self-preservation especially successful, those involved may have to "do something extra", without sufficient evidence for obtaining return benefits. In particular, it would be unwise to immediately stop interacting after observing that the other stays behind in contributing to a common goal. It may even be important to start behaving nicely to others without *any* evidence that the other will be nice too. It is argued here that properties such as patience, tolerance, and forgiveness are associated with a strongly developed care system and that such a system provides a plausible mechanistic explanation for a tit-for-tat strategy in a repeated prisoner' dilemma (Axelrod and Hamilton 1981).

Finally, if individuals are capable of symbolically representing the social interdependencies involved in cooperation, the care mechanism as sketched above may generate a wide variety of moral emotions that help to sustain cooperation during "critical moments", when there is both temptation and opportunity to cheat. This may be what research with the prisoner's dilemma and other social dilemma's actually is about: Human subjects capable of seeing and mentally representing the complete payoff matrix of the particular game as a *symbolic* representation of social interdependencies and vulnerabilities, and deriving their game behavior from a variety of emotional reactions triggered by this representation. (It seems doubtful if it makes sense to assume that species without symbolic capacity can play this game in a similar sense.) In particular, mentally representing the payoff matrix of the prisoner's dilemma implies that players are aware that both the self and the other player are dependent on each other (and find themselves in a vulnerable position), especially in the sense that both may be tempted to defect because this is associated with the greatest individual benefit. Especially in making the first move in the prisoner's dilemma game, one may more or less trust that the other will care about one's vulnerable position (for an interpretation of trust in terms of vulnerability, see Rousseau et al. 1998).

Now, causing or imagining causing harm on vulnerable beings (which are likely to activate the care system) is responsible for a variety of moral emotions. Thus when trust is broken, the other player's defection and its resulting unequal distribution of valuable goods, may arouse moral anger, whereas defection by the self may induce guilt. Importantly, in imagining or anticipating these outcomes, similar emotions may be aroused, of which especially guilt would be an important motivator to *prevent* causing harm on the vulnerable other (Nelissen et al. 2007). In case of the other's defection, pity or forgiveness may also be aroused, especially if the defection can be attributed to external causes beyond the individual's control such as illness (Batson and Ahmad 2001) or noise (Van Lange et al. 2002). The fact that nasal administration of oxytocin, the hormone which is typically associated with activity of the parental care system (Panksepp 1998; Uvnäs-Moberg 1998), increases trust and fair behavior in public good games that are related to the prisoner's dilemma (Kosfeld et al. 2005), also seems to support an interpretation of cooperative behavior in terms of an activated (parental) care mechanism (on neurophysiological evidence for the presence of a parental care system in mammals, including humans, and its interaction and competition with a self-preservational fight-or-flight system, see Kirsch et al. 2005; Panksepp 1998; Uvnäs-Moberg 1998). Involvement of the parental care mechanism is further suggested by demonstrations that cooperation in the prisoner's dilemma or similar social dilemma's is positively influenced by exposure to physical cues associated with infants and care such as a smile or touch (e.g., Boone and Buck 2003; Mehu et al. 2007).

Similar explanations may be used to understand behavior in the ultimatum game, often used by economists to study other-benefiting, in which one person is asked to divide a sum of money (supplied by the experimenter) between him or her and another person (for a review, see Sigmund et al. 2002). It has been observed in many different cultures that the majority of players behave in a costly manner; they share close to 50%, and do not accept offers lower than 20%. In terms of the present theory this behavior can be explained by assuming that money supplied to one of the players is not "owned" yet by anyone of the players, which means that the one who is allowed to distribute it may imagine to cause a relative need state in the (vulnerable) other player by not giving half of it to him or her; subsequently resulting in moral emotions such as sympathy or guilt.

The Evolution of a Uniquely Human Capacity for Cooperation

In the previous sections, it was suggested that humans are endowed with a psychological mechanism that not only

strongly motivates them to engage in asymmetrical helping of needy individuals but also provides a background "softness" or gentleness that is needed for social properties that can be described as tolerance, patience, forgiveness, doing something extra, and moral emotions, and which seem essential to make symmetrical cooperation among highly autonomous individuals a success. Yet, this softness may be less visible in hierarchically organized societies or under conditions in which self-preservational motives are strongly activated. It was proposed that the underlying mechanism bears all the marks of a mammalian parental care system acting together with self-preservational mechanisms, yet integrated with a capacity for symbolic representation and reasoning. This section briefly speculates about the evolutionary processes in humans that have led to its generalization beyond parent-offspring relationships.

First note that especially theorists adopting the prisoner's dilemma as the main paradigm for explaining the evolution of cooperation (e.g., Cosmides and Tooby 1992; Trivers 1971), would argue that a mechanism with a relatively unconditional and indiscriminate care element may not evolve at all in an environment in which at least some individuals are likely to misuse it and cheat. In responding to this objection, it is important to recognize that a social relationship or group to a certain extent *can* bear the occurrence of social parasitism or cheating. Although the psychological mechanism proposed in this paper automatically generates unconditional care when needed, it also allows individuals to be critical and selective toward others and to take into account self-preservational concerns. Social parasites will first be punished in the relatively tolerant and soft manner typical for mammalian parents, yet will receive more aggression as evidence increases that they have matured sufficiently and can stand on their own feet (see Trivers 1974, on parent-offspring conflict). Such a view is consistent with Axelrod and Hamilton' (1981) interpretation of the emergence of cooperation among nonkin in the prisoner's dilemma game, arguing that it can generalize to nonkin after it has first been firmly established among kin.

The following reasons may be mentioned for expecting the genes for an exceptionally strong or easily activated parental care mechanism to be generally present in the human population. First, bipedality in humans is associated with a reduced size of the birth channel, and hence giving birth to altricial babies that require an exceptionally strong maternal motivation to engage in protection and nurturance, sustained during an extraordinarily long period of immaturity.

Second, despite the extensive care and delays in reproduction demanded by human infants, inclusive fitness may have increased by human parents starting to select or "breed" offspring in which mutations caused the parental care mechanism to get active before sexual maturity; thus also allowing immature siblings to respond with care to each others' needs and suffering (for a similar suggestion, see Dawkins 1976/1989, p. 281). This would free parents to engage in other activities associated with self-preservation and reproduction and would result in mature offspring that would be well prepared for providing similarly effective long-term care to their own offspring, and so on. The presence of an easily activated parental care mechanism in human children is suggested by children's nurturing and helping responses to vulnerable ("cute") things such as younger siblings, animals, and dolls (Fogel et al. 1986; Hrdy 2009; Warneken and Tomasello 2009), and the occurrence of moral emotions and corresponding behavior in early childhood (Eisenberg 2000). Importantly, human infants also show relatively unconditional and indiscriminate helping tendencies toward strange adults in need, and only later start to discriminate on the basis of group membership and reciprocity (Warneken and Tomasello 2009). That a mechanism for parental care may be involved in helping among adults is suggested by the fact that the cues that can trigger prosocial behavior often tend to be associated with the craniofacial features of neonates and toddlers, even when present in adult faces (Berry and McArthur 1986; Keating et al. 2003; Lorenz 1943).

The result of generally present genes for a parental care mechanism would be a society in which all members have the psychological capacity (in some individuals more strongly developed than in others, and sensitive to contextual and cultural influences) for both asymmetrical helping and symmetrical cooperation. The present theory also suggests a motivational and moral mechanism for cooperative breeding, which until now has been largely interpreted in terms of cognitive aspects (Burkart and van Schaik 2010). Perhaps, group selection plays an additional role in the evolution of relatively indiscriminate altruism toward ingroup members (e.g., Sober and Wilson 1998).

Finally note how a combination of a strongly activated parental care mechanism and uniquely human intelligence complements the *social brain hypothesis* (Dunbar 2003; Humphrey 1976) which states that, in primates, brain size and associated intelligence have primarily increased in the service of "Machiavellian intelligence". This kind of intelligence may indeed be characteristic for a hierarchically organized and perhaps male-dominated society or under strong competition. Yet, caregivers, whether they be parents, young children, or cooperating strangers, must employ a different kind of intelligence also requiring an increased brain size; one that allows them to engage effectively in the many tasks necessary to care effectively for vulnerable others.

Conclusions

The physical and mechanistic constraints on social behavior uncovered in this paper cannot easily come to the fore when primarily adopting a game-theoretical perspective on social interaction. As argued, these constraints reveal themselves first when it is recognized that, in order to produce adaptive behavior, behavioral mechanisms in general have to be based on accurate perception and internal representation of the physical properties of other individuals and the environment in which they live. Some species manage to respond adaptively primarily on the basis of reactive mechanisms, others may complement reactivity with accurate internal representation of, and even reasoning about, invariant properties.

The most important results of this mechanistic analysis can be summarized as follows. There are two broad classes of behaviors involved in providing benefits to others: Helping and cooperation. Adopting this classification has allowed us to identify a specifically designed psychological mechanism for effective asymmetrical other-benefiting, explaining in a mechanistic sense the existence of "unconditional kindness" and a variety of moral emotions in response to perceived vulnerability. This classification has also allowed us to distinguish a limited number of mechanisms associated with self-preservation that are involved in cooperation. Yet, this distinction has also made it possible to recognize that the mechanisms involved in helping and cooperation may act together; in such a way that highly autonomous agents such as humans experience sufficient motivation to conscientiously continue to cooperate during critical moments.

In light of a game-theoretical perspective, where choosing a particular row or column of a payoff matrix has unclear physical correlates, it is difficult to obtain these insights; a marriage between game theory and traditional cognitive psychology makes it even more difficult. In particular, an emphasis on symbolic representation and reasoning makes it difficult to account for properties of psychological mechanisms such as reactivity; variation in intensity of sensory stimulation, motivation, and motor output; competition and mutual inhibition between different mechanisms; and accurate internal representation of object properties in terms of sensorimotor networks without the use of symbols.

This present critique of game theory, however, is not meant to replace it. One interesting connection between game theory and the present approach is that a game and its payoff matrix can be conceived as a symbolic representation of social interdependencies that can be mentally represented by certain species. In this way, one can study how emotional reactions to such a mental representation and conscience in general can motivate players to make cooperative choices. Interestingly, the use of games by psychologists and economists once in a while reveals how the presently proposed mechanism for other-benefiting may be involved in actual game playing (see the effects of certain nonverbal expressions and oxytocin on cooperative choices, discussed earlier).

In light of the present emphasis on physical and mechanistic constraints on the evolution of cooperation, it may be proposed that multi-agent modeling rather than game theory is currently the most promising research paradigm for studying it. In particular, multi-agent modeling allows researchers to endow agents in a realistic way with multiple psychological mechanisms that are triggered by specific features and behaviors of other agents or their common environment, and to vary the activation thresholds of these mechanisms (Epstein and Axtell 1996). A relevant example would be a simulation by Cesta, Miceli, and Rizzo (1996), showing that simple agents that unconditionally help others that beg for food and parasitize on helpful others (i.e., do not look for food themselves when needy), can survive and remain phenotypically dominant in a population as long as they do not provide help when their own need for food exceeds a critical value. These phenomena cannot be made visible and explained by treating organisms as nonphysical players in a game-theoretical approach to cooperation. They remind us that evolutionary explanations for other-benefiting behavior should respect physical limitations and possibilities in designing and constructing adaptive behavioral mechanisms.

References

Altemeyer, B. (1998). The other "authoritarian personality". In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 30, pp. 47–92). Orlando, FL: Academic Press.

Alvard, M. S. (2001). Mutualistic hunting. In C. Stanford & H. Bunn (Eds.), *The early human diet: The role of meat* (pp. 261–278). New York: Oxford University Press.

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

Batson, C. D. (1987). Prosocial motivation: Is it ever truly altruistic? In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 20, pp. 65–122). New York: Academic Press.

Batson, C. D. (1998). Altruism and prosocial behavior. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 2, pp. 282–316). New York: McGraw-Hill.

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

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanistic alternative. *Studies in the History and Philosophy of Biological and Biomedical Sciences, 36*, 421–441.

Bergmüller, R., Johnstone, R. A., Russell, A. F., & Bshary, R. (2007). Integrating cooperative breeding into theoretical concepts of cooperation. *Behavioural Processes, 76*, 61–72.

Berry, D. S., & McArthur, L. Z. (1986). Perceiving character in faces: The impact of age-related craniofacial changes on social perception. *Psychological Bulletin, 100*, 3–18.

Bindra, D. (1985). Motivation, the brain, and psychological theory. In S. Koch & D. E. Leary (Eds.), *A century of psychology as science* (pp. 338–363). New York: McGraw-Hill.

Binmore, K. (2007a). *Game theory: A very short introduction*. Oxford: Oxford University Press.

Binmore, K. (2007b). *Playing for real: A text on game theory*. New York: Oxford University Press.

Boehm, C. (1999). *Hierarchy in the forest: The evolution of egalitarian behavior*. Cambridge, MA: Harvard University Press.

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

Brooks, R. A. (1997). Intelligence without representation. In J. Haugeland (Ed.), *Mind design II: Philosophy, psychology, artificial intelligence* (pp. 395–420). Cambridge, MA: MIT Press.

Brosnan, S. F., Salwiczek, L., & Bshary, R. (2010). The interplay of cognition and cooperation. *Philosophical Transactions of the Royal Society of London B, 365*, 2699–2710.

Bryson, J. (2000). Cross-paradigm analysis of autonomous agent architecture. *Journal of Experimental and Theoretical Artificial Intelligence, 12*, 165–189.

Bshary, R., & Bergmüller, R. (2008). Distinguishing four fundamental approaches to the evolution of helping. *Journal of Evolutionary Biology, 21*, 405–420.

Bunge, M., & Ardilla, R. (1987). *Philosophy of psychology*. New York: Springer.

Burkart, J. M., & van Schaik, C. P. (2010). Cognitive consequences of cooperative breeding in primates? *Animal Cognition, 13*, 1–19.

Cacioppo, J. T., & Berntson, G. G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin, 115*, 401–423.

Caporael, L., Dawes, R. M., Orbell, J. M., & van de Kragt, A. J. C. (1989). Selfishness examined: Cooperation in the absence of egoistic incentives. *Behavioral and Brain Sciences, 12*, 683–699.

Cesta, A., Miceli, M., & Rizzo, P. (1996). Help under risky conditions: Robustness of the social attitude and system performance. In V. Lesser (Ed.), *Proceedings of the second international conference on multiagent systems* (pp. 18–25). Menlo Park, CA: AAAI Press.

Clark, A. (1999). An embodied cognitive science. *Trends in Cognitive Sciences, 3*, 345–351.

Clutton-Brock, T. H. (2009). Cooperation between non-kin in animal societies. *Nature, 462*, 51–57.

Clutton-Brock, T. H., & Parker, G. A. (1995). Punishment in animal societies. *Nature, 373*, 209–216.

Connor, R. C. (1995). The benefits of mutualism: A conceptual framework. *Biological Reviews, 70*, 427–457.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 163–228). New York: Oxford University Press.

Cosmides, L., & Tooby, J. (2005). Neurocognitive adaptations designed for social exchange. In D. M. Buss (Ed.), *The handbook of evolutionary psychology* (pp. 584–626). Hoboken, NJ: Wiley.

Costa, P., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology, 81*, 322–331.

Darwin, C. (1872/1998). *The expression of the emotions in man and animals (with an introduction, afterword, and commentaries by Paul Ekman)* (3 ed.). London: Harper Collins.

Dawkins, R. (1976/1989). *The selfish gene (new edition)*. Oxford: Oxford University Press.

Dawkins, R., & Krebs, J. R. (1979). Arms races within and between species. *Proceedings of the Royal Society London B, 205*, 412–480.

de Waal, F. B. M. (1996). *Good natured: The origins of right and wrong in humans and other animals*. Cambridge, MA: Harvard University Press.

de Waal, F. B. M. (2008). Putting the altruism back into altruism: The evolution of empathy. *Annual Review of Psychology, 59*, 279–300.

de Waal, F. B. M., & Suchak, M. (2010). Prosocial primates: selfish and unselfish motivations. *Philosophical Transactions of the Royal Society of London B, 365*, 2711–2722.

Dijker, A. J. M. (2001). The influence of perceived suffering and vulnerability on the experience of pity. *European Journal of Social Psychology, 31*, 659–676.

Dijker, A. J. M. (2008). Why Barbie feels heavier than Ken: The influence of size-based expectancies and social cues on the illusory perception of weight. *Cognition, 106*, 1109–1125.

Dijker, A. J. M. (2010). Perceived vulnerability as a common basis of moral emotions. *British Journal of Social Psychology, 49*, 415–423.

Dijker, A. J. M., & Koomen, W. (2007). *Stigmatization, tolerance, and repair: An integrative psychological analysis of responses to deviance*. Cambridge: Cambridge University Press.

Dunbar, R. I. M. (2003). The social brain: Mind, language, and society in evolutionary perspective. *Annual Review of Anthropology, 32*, 163–181.

Eisenberg, N. (2000). Emotion, regulation, and moral development. *Annual Review of Psychology, 51*, 665–696.

Epstein, J. M., & Axtell, R. (1996). *Growing artificial societies: Social science from the bottom up*. Cambridge, MA: MIT Press.

Fábrega, H. (1997). *Evolution of sickness and healing*. Berkeley, CA: University of California Press.

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review, 90*, 980–994.

Fogel, A., Melson, G. F., & Mistry, J. (1986). Conceptualizing the determinants of nurturance: A reassessment of sex differences. In A. Fogel & G. F. Melson (Eds.), *Origins of nurturance: Developmental, biological and cultural perspectives on caregiving* (pp. 53–67). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gardner, A., & Foster, K. R. (2008). The evolution and ecology of cooperation: History and concepts. In J. Korb & J. Heinze (Eds.), *Ecology of social evolution* (pp. 1–36). Berlin: Springer.

Gault, B. A., & Sabini, J. (2000). The roles of empathy, anger, and gender in predicting attitudes toward punitive, reparative, and preventative public policies. *Cognition and Emotion, 14*, 495–520.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior, 24*, 153–172.

Gorniak, S. L., Zatsiorsky, V. M., & Latash, M. L. (2009). Manipulation of a fragile object. *Experimental Brain Research, 193*, 615–631.

Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review, 25*, 161–178.

Gregory, R. L. (2005). Knowledge for vision: Vision for knowledge (the Medawar lecture 2001). *Philosophical Transactions of the Royal Society B, 360*, 1231–1251.

Grossberg, S. (1980). How does the brain build a cognitive code? *Psychological Review, 87*, 1–51.

Hagen, E. H., & Hammerstein, P. (2006). Game theory and human evolution: A critique of some recent interpretations of experimental games. *Theoretical Population Biology, 69*, 339–348.

Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective science* (pp. 852–870). Oxford: Oxford University Press.

Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316*, 998–1002.

Hamilton, W. D. (1964). The genetical evolution of social behaviour (I and II). *Journal of Theoretical Biology, 7*, 1–52.

Hamilton, W. D. (1971). Geometry for the selfish herd. *Journal of Theoretical Biology, 31*, 295–311.

Hammerstein, P. (2003a). Why is reciprocity so rare in social animals? A protestant appeal. In P. Hammerstein (Ed.), *Genetic and cultural evolution of cooperation* (pp. 83–93). Cambridge, MA: MIT Press.

Hammerstein, P. (Ed.). (2003b). *Genetic and cultural evolution of cooperation*. Cambridge, MA: MIT Press.

Helmholtz, H. L. (1878). The facts of perception. In R. Kahl (Ed.), *Selected writings of Herman von Helmholtz*. Middletown, Conn: Wesleyan University Press.

Hofstede, G. (2001). *Culture's consequences: comparing values, behaviors, institutions, and organizations across nations*. Thousand Oaks: Sage Publications.

Hölldobler, B., & Wilson, E. O. (1990). *The ants*. Harvard: Harvard University Press.

Hrdy, S. B. (2009). *Mothers and others: The evolutionary origins of mutual understanding*. Cambridge, MA: Harvard University Press.

Humphrey, N. K. (1976). The social function of intellect. In P. P. G. Bateson & R. A. Hinde (Eds.), *Growing points in ethology* (pp. 303–317). Cambridge: Cambridge University Press.

Inglehart, R., & Baker, W. E. (2000). Modernization, cultural change, and the persistence of traditional values. *American Sociological Review, 65*, 19–51.

Jaffee, S., & Hyde, J. S. (2000). Gender differences in moral orientation: A meta-analysis. *Psychological Bulletin, 126*, 703–726.

Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain Sciences, 17*(2), 187–245.

Keating, C. F., Randall, D. W., Kendrick, T., & Gutshall, K. A. (2003). Do babyfaced adults receive more help? The (cross-cultural) case of the lost resume. *Journal of Nonverbal Behavior, 27*, 89–109.

Kilner, R., & Johnstone, R. A. (1997). Begging the question: Are offspring solicitation behaviours signals of need? *Trends in Ecology & Evolution, 12*, 11–15.

Kirsch, P., Esslinger, C., Chen, Q., Mier, D., Lis, S., Siddhanti, S., et al. (2005). Oxytocin modulates neural circuitry for social cognition and fear in humans. *The Journal of Neuroscience, 25*, 11489–11493.

Knauft, B. M. (1991). Violence and sociality in human evolution. *Current Anthropology, 32*, 391–428.

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

Korchmaros, J. D., & Kenny, D. A. (2006). An evolutionary and close-relationship model of helping. *Journal of social and Personal Relationships, 23*, 21–43.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature, 435*, 673–676.

Lehmann, L., & Keller, L. (2006). The evolution of cooperation and altruism: A general framework and classification of models. *Journal of Evolutionary Biology, 19*, 1365–1725.

Leimar, O., & Hammerstein, P. (2010). Cooperation for direct fitness benefits. *Philosophical Transactions of the Royal Society B, 365*, 2619–2626.

Lewis, M. D. (2005). Bridging emotion theory and neurobiology through dynamic systems modeling. *Behavioral and Brain Sciences, 28*, 169–245.

Lorenz, K. (1943). Die angeborenen Formen möglicher Erfahrung [The innate forms of potential experience]. *Zeitschrift für Tierpsychologie, 5*, 235–409.

Ludlow, A. R. (1980). The evolution and simulation of a decision maker. In F. M. Toates & T. R. Halliday (Eds.), *Analysis of motivational processes*. London: Academic Press.

Maestripieri, D. (1999). The biology of human parenting: Insights from nonhuman primates. *Neuroscience and Biobehavioral Reviews, 23*, 411–422.

Maynard Smith, J., & Szathmary, E. (1995). *The major transitions in evolution*. Oxford: W. H. Freeman.

McCullough, M. E., Kilpatrick, S. D., Emmons, R. A., & Larson, D. B. (2001). Is gratitude a moral affect? *Psychological Bulletin, 127*, 249–266.

McDougall, W. (1908/1948). *An introduction to social psychology* (29th ed.). London: Methuen.

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

Miller, N., Pedersen, W. C., Earleywine, M., & Pollock, V. E. (2003). A theoretical model of triggered displaced aggression. *Personality and Social Psychology Review, 7*, 75–97.

Nelissen, R. M. A., Dijker, A. J. M., & de Vries, N. K. (2007). How to turn a hawk into a dove and vice versa: Interactions between emotions and goals in a give-some dilemma game. *Journal of Experimental Social Psychology, 43*, 280–286.

Nelissen, R. M. A., & Zeelenberg, M. (2009). When guilt evokes self-punishment: Evidence for the existence of a *Dobby effect*. *Emotion, 9*, 118–122.

Noë, R. (2006). Cooperation experiments: Coordination through communication vs. acting apart together. *Animal Behaviour, 71*, 1–18.

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

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*, 939–1031.

Okasha, S. (2010). Altruism researchers must cooperate. *Nature, 467*, 653–655.

Packer, C., & Ruttan, R. (1988). The evolution of cooperative hunting. *The American Naturalist, 132*, 159–198.

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotion*. New York: Oxford University Press.

Parsons, T. (1951). *The social system*. New York: The Free Press.

Penner, L. A., Dovidio, J. F., Piliavin, J. A., & Schroeder, D. A. (2005). Prosocial behavior: Multilevel perspectives. *Annual Review of Psychology, 56*, 365–392.

Preston, S. D., & de Waal, F. B. M. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences, 25*, 1–20.

Richerson, P. J., & Boyd, R. (2005). *Not by genes alone: How culture transformed human evolution*. Chicago: University of Chicago Press.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review, 23*, 393–404.

Shepard, R. N. (1989). Internal representation of universal regularities: A challenge for connectionism. In L. Nadel, L. Cooper, P. Culicover, & R. M. Harnish (Eds.), *Neural connections,*

mental computation: Computational models of cognition and perception (pp. 104–134). Cambridge, MA: MIT Press.

Sigmund, K., Fehr, E., & Nowak, M. A. (2002). The economics of fair play. *Scientific American*, *83*, 83–87.

Silk, J. B. (2003). Cooperation without counting: The puzzle of friendship. In P. Hammerstein (Ed.), *Genetic and cultural evolution of cooperation* (pp. 37–54). Cambridge, MA: MIT.

Skoe, E. E. A., Eisenberg, N., & Cumberland, A. (2002). The role of reported emotion in real-life and hypothetical moral dilemmas. *Personality and Social Psychology Bulletin, 28*, 962–973.

Sloman, A., & Chrisley, R. (2005). More things than are dreamt of in your biology: Information processing in biologically-inspired robots. *Cognitive Systems Research, 6*, 145–174.

Smith, A. (1776/1910). *The wealth of nations*. London: J. M. Dent & Sons.

Sober, E., & Wilson, D. S. (1998). *Unto others: The evolution and psychology of unselfish behavior*. Cambridge, MA: Harvard University Press.

Sommerhoff, G. (1974). *Logic of the living brain*. London: Wiley.

Staats, A. W. (1968). Social behaviorism and human motivation: Principles of the Attitude-reinforcement-discrimination system. In A. G. Greenwald, T. C. Brock, & T. M. Ostrom (Eds.), *Psychological foundations of attitudes*. San Diego, CA: Academic Press.

Sterelny, K. (1990). *The representational theory of mind: An introduction*. Oxford, GB: Basil Blackwell.

Stevens, J. R., Cushman, F. A., & Hauser, M. D. (2005). Evolving the psychological mechanisms for cooperation. *Annual Review of Ecology, Evolution, and Systematics, 36*, 499–518.

Stijnen, M. M. N., & Dijker, A. J. M. (in press). Reciprocity and need in posthumous organ donation: The mediating role of moral emotions. *Social Psychological and Personality Science*.

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology, 58*, 345–372.

Toates, F. (1986). *Motivational systems*. Cambridge: Cambridge University Press.

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19–136). New York: Oxford University Press.

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

Trivers, R. L. (1974). Parent-offspring conflict. *American Zoologist, 14*, 249–264.

Trivers, R. L. (1985). *Social evolution*. Menlo Park, CA: Benjamin/Cummings.

Tuomela, R. (2000). *Cooperation*. Dordrecht, The Netherlands: Kluwer.

Uvnäs-Moberg, K. (1998). Oxytocin may mediate the benefits of positive social interaction and emotions. *Psychoneuroendocrinology, 23*, 819–835.

Van Lange, P. A. M., Ouwekerk, J. W., & Tazelaar, M. J. A. (2002). How to overcome the detrimental effects of noise in social interaction: The benefits of generosity. *Journal of Personality and Social Psychology, 82*, 768–780.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton: Princeton University Press.

Warneken, F., & Tomasello, M. (2009). The roots of human altruism. *British Journal of Psychology, 100*, 455–471.

West, S. A., Griffin, A. S., & Gardner, A. (2007). Social semantics: Altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology, 20*, 415–432.

Wilson, E. O. (1980). *Sociobiology: The abridged edition*. Cambridge, MA: Belknap Press of Harvard University Press.

Wispé, L. (1991). *The psychology of sympathy*. New York: Plenum.

Worden, L., & Levin, S. A. (2007). Evolutionary escape from the prisoner's dilemma. *Journal of Theoretical Biology, 245*, 411–422.