Journal of Neuroscience, Psychology, and Economics
2014, Vol. 7, No. 1, 33–47

1937-321X/14/$12.00 DOI: 10.1037/npe0000015

Caring, the Emotions, and Social Norm Compliance

Matteo Colombo
Tilburg University

Does emotion motivate social norm compliance? This paper addresses this question by examining some of the aspects of the motivational structure of social norm compliance. It argues two claims: First, neither resentment nor pleasure is the ultimate motive of norm compliance; second, caring is necessary, both to feel certain emotions and to comply with norms. According to my proposal, caring is a relatively stable volitional profile, something through which agents steadily orient their behaviors within their environments in pursuit of a good life. Explicating this notion, at both the personal and subpersonal level of explanation, helps to see in what sense we cannot feel certain emotions or comply with social norms unless we possess the capacity to care.

Keywords: caring, social norms, resentment, pleasure, meuromodulators

There is little doubt that emotion plays an important role in the regulation of our moral and social lives. Yet, it is controversial how it is, exactly, that emotion motivates people to abide by social norms. The empirical evidence from psychology and the brain sciences does not warrant firm conclusions (for concise reviews, see Huebner, Dwyer, & Hauser, 2009; Montague & Lohrenz, 2007), and the philosophical debate has focused either on the relationship between emotion and norm violation, or between emotion and normative judgment (e.g., Sinnott-Armstrong, 2008a, 2008b). This paper contributes to existing literature in moral psychology and social decision making by examining some of the aspects of the motivational structure of social norm compliance. It examines two prominent arguments: Robert Sugden's (1998, 2000) resentment hypothesis, and Ernst Fehr and Colin Camerer's (2007) hedonist hypothesis of norm compliance. Two claims are defended: First, neither resentment nor pleasure is the ultimate motive for norm compliance; second, caring is necessary both to feel certain emotions and to comply with social norms.

The paper is organized thus. Section 1 engages one of the few explicit theoretical arguments that emotion is the ultimate motive for norm compliance: Robert Sugden's resentment hypothesis (Sugden, 1998, 2000). Using Sugden's argument as a foil advances the points made in this paper—which does not hinge on any sophisticated account of emotion—because Sugden assumes a commonsensical view of emotions understood as feelings that people experience. I argue, contra Sugden, that resentment is not the motivational source of norm compliance. In doing so, I single out the capacity to care as necessary to feel at least some emotions. Section 2 focuses on the positive emotion of pleasure. It argues that Fehr and Camerer's (2007) hedonistic interpretation of neurobiological data about social norm compliance is unjustified. Section 3 sharpens the notion of caring by characterizing some core components of its structural and neurophysiological basis. The conclusion summarizes the original contribution of the paper to existing literature and formulates some questions for further research.

This article was published Online First November 25, 2013.

This work was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program New Frameworks of Rationality (SPP 1516).

I am sincerely grateful to Courtney Humeny, Claudio Salvatore, Koosha Eghbal Ketabchi, and Ryan Muldoon for helpful discussions on some of the ideas contained here. A special thank you goes to Daniel Houser and an anonymous reviewer for their constructive criticisms, and suggestions.

Correspondence concerning this article should be addressed to Matteo Colombo, Tilburg Center for Logic and Philosophy of Science, Tilburg University, Post Office Box 90153, 5000 LE Tilburg, the Netherlands. E-mail: m.colombo@uvt.nl

1 Resentment and Norm Compliance

Imagine you are traveling on a crowded train without a seat. You are tired of standing, and someone leaves her seat to go to the toilet. Why don't you take her seat? A plausible explanation may invoke the existence of a norm that bounds the set of appropriate actions in that type of context. In the vocabulary of folk psychology: You don't take the seat and you keep standing because you believe that taking the seat of someone who leaves it to go to the toilet falls outside the relevant set of appropriate actions, and you find that social norm to be reasonable.

Robert Sugden (1998, 2000) argues that it is not the acceptance of the norm that plays a fundamental role in motivating compliance with it. Sugden develops an "emotional sanctioning" account of norm compliance. He aims to explain the emergence of social norms in general, and norm compliance in particular, with no appeal to normative concepts. The core of Sugden's argument is an empirical hypothesis called the resentment hypothesis, which provides sufficient conditions for the arousal of resentment. Resentment, for Sugden, is a nonmoral sentiment; it does not depend on any moral code. What ultimately motivates norm compliance would be the feeling of resentment.

Before examining the resentment hypothesis, let me clarify how "emotion" is used here. The term "emotion" is used in different ways by different researchers to refer to distinct psychological phenomena (cf. Adolphs, 2010; de Sousa, 2010). Sugden uses "emotion" interchangeably with "sensation," "sentiment," "affect," and "feeling." He is influenced by Adam Smith's (1759/1976) theory of moral sentiments (cf. Sugden, 2002). Smith provides a commonsensical account of various feelings such as resentment and sympathy, which we are invited to test against our own experience. Accordingly, I use "emotion" in an ordinary sense, as a type of feeling (cf. Bennett & Hacker, 2003, Ch. 7).

Sugden begins by claiming that when other people's actions constitute a predictable behavioral pattern, they thereby impose "some obligation on me to conform to that pattern" (Sugden, 2000, p. 112). The claim is not about the existence of a general moral principle. It is not that there exists some obligation to conform to behavioral patterns in virtue of their being predictable. The claim is that "people are in fact motivated as if by some such principle" (Sugden, 2000, p. 112). Hence, the aim here is to explain or predict people's social norm compliance. Sugden is *not* arguing that certain normative considerations should ground the motivation to comply with social norms. His argument is descriptive: Certain behavioral patterns are associated with particular "normative expectations," which motivate us to comply with norms, courtesy of specific affective signatures. When one has a normative expectation, he or she expects that others expect him or her to do something. But how is it that the fact that some people expect one to do Φ in a certain type of situation (*S*) makes him or her want to do Φ in *S*?

According to Sugden, we naturally feel resentment against those who act contrary to our expectations and we feel aversion toward frustrating others' expectations. By "resentment," Sugden means "a sensation or sentiment that compounds disappointment at the frustration of one's expectations with anger and hostility directed at the person who is frustrating (or has frustrated) them" (Sugden, 2000, p. 113). Aversion depends on resentment. A person conforms to a behavioral pattern because others will resent him or her otherwise; that person knows this and is emotionally averse to others' resentment. Resentment and aversion may be intertwined with cognitive states and processes. Here, "cognitive" qualifies processes and states, such as knowledge and belief, which contrast with affections and emotions.

There are two ways in which cognitive states and processes enter Sugden's account of norm compliance. First, sometimes people feel resentment, and, at the same time, have knowledge that they have been wronged, given some normative standard. However, people do not feel resentment because of their normative knowledge. That person *j* feels resentment at person *i*'s doing Φ doesn't presuppose that *j* believes that *i* ought not to Φ. For example, your friend and you have agreed to meet for lunch. You are waiting for her, when she phones you telling you that she is ill and she cannot make it. Although you know that your feeling is unjustified, you may feel resentment toward your friend in this situation. Similarly, that person *i* feels aversion toward doing Φ doesn't

presuppose any belief by *j* that he or she ought not to do Φ.

For Sugden, resentment and aversion are more fundamental than ought-beliefs in two ways. On the one hand, resentment and aversion are evolutionarily more primitive than cognitive states such as ought-beliefs. On the other, many of our ought-beliefs "are nothing more than generalizations of more primitive sentiments" like resentment and aversion (Sugden, 2000, p. 115). When ought-beliefs motivate people to comply with norms in particular cases, they do so in virtue of resentment and aversion, of which they are generalizations. Hence, resentment and aversion are also more fundamental than ought-beliefs in motivating norm compliance in particular cases.

There is a second way in which cognitive states are linked to resentment and aversion. This leads to the formulation of the resentment hypothesis, which relies on common knowledge conditions. Specifically

Let *P* be a population and *I* a behavioral pattern dependent on some interaction among the individuals in *P*. Let *i* and *j* be any two individuals from *P* that engage in *I*. Let Φ and Ψ be alternative actions that *i* can take in situation *S*. Whichever action *i* decides to take, it will be common knowledge after the event. Assume that it is common knowledge within *P* that individuals in *i*'s position normally do Φ rather than Ψ. It is also common knowledge within *P* that people in *j*'s position have grounds to expect *i* to do Φ and that they normally prefer that people in *i*'s position do Φ rather than Ψ. Granted that *j* has that preference, then *i*'s doing Ψ will induce in *j* a feeling of resentment toward *i*; *i*'s awareness of this will induce in *i* a feeling of aversion toward doing Ψ (Sugden, 2000, pp. 114–116).

The resentment hypothesis predicts that people will feel resentment toward those who fail to conform to their expectations. Because this tendency of people to feel resentment is common knowledge, people will tend to avoid acting in ways that likely provoke feelings of resentment. The hypothesis is stated as a sufficient condition for the arousal of resentment. The basic idea is that a "person can be motivated to meet other people's expectations about him" and this motivation is grounded in an emotion (Sugden, 2000, p. 116).

Sugden illustrates how the sentiment of resentment motivates norm compliance with the following type of example. It is well-known that diners in the United States leave tips of at least 15% of the bill. I know this fact. I have good reason to expect that waitresses in the U.S. expect me to leave a 15% tip if I dine out in the U.S. I go to a restaurant in the U.S., but I am Italian and I have little interest in meeting the waitress's expectation. Still, the existence of the expectation will motivate me to tip her. If I don't tip, I will feel uneasy and embarrassed. I am emotionally averse to those emotions, and this aversion motivates me to comply with the norm of tipping.

1.1 Not by Resentment Alone

I believe that Sugden's hypothesis is not sufficient. My claim is that the resentment hypothesis is plausible only within a population whose people care about each other's preferences, expectations, and behaviors. This notion of caring will be explicated first by ostension, by pointing to the relevant phenomenon with a number of cases. In the following subsection, I draw on these cases to elucidate what "caring" can plausibly mean.

The argument developed in this section can be summarized thus:

P1. The resentment hypothesis depends on an individual, *j*, preferring agent *i* doing Φ.
P2. If an individual *j* feels resentment about agent *i* doing Φ, then *j* cares about *i* doing Φ.
P3. Caring is distinct from preferring.
P4. Sometimes an individual *j* prefers things about which she or he doesn't care.
P4′. Sometimes *j* prefers *i* to do Φ while *j* doesn't care about *i* doing Φ.
C1. Sometimes *j* doesn't feel resentment that *i* doesn't do Φ, even if *j* prefers *i* do Φ.
C2. The resentment hypothesis is in general insufficient.

P1 describes one of the conditions in the resentment hypothesis. P2 claims that caring about something is necessary for feeling emotions about it. We should distinguish between two issues: under what conditions we feel resentment, and under what conditions we are

affected by other people's resentment toward us. P2 can then be understood as making two claims: We feel emotions only for people, objects, behavioral patterns we care about; we are emotionally affected by other people's resentment toward us only if we care about what other people feel, prefer, or think about us. P3 and P4 are related. P4′ is a special case of P4. C1 and C2 follow from the five premises.

I start to argue for P2 by building on Elizabeth Anderson's (2000) response to Sugden's (2000) account, which, she argues, is incoherent. She focuses on the conditions under which people's decision to comply with norms is affected by others' resentment toward them. Sugden—she reasons—assumes that people can feel resentment on behalf of others, because we all share the same basic nonmoral sentiments. But then norm violators should resent themselves. They need not be averse to others' resentment to be motivated to comply with norms. "Given the impartiality of moral sentiments, they can just as easily be directed against [themselves] as against any other person" (Anderson, 2000; p. 184). Hence, other people's normative expectations could well be superfluous in motivating one to comply with norms. If self-resentment can be enough for norm compliance, then an individual can care about complying (or not complying) with social norms independently of what others expect. In other words, people can have an intrinsic motivation to comply with norms: They can "comply with norms as ultimate ends, rather than as a means to other ends" (Sripada & Stich, 2007, p. 281).

A criticism of Anderson's argument is that, in general, the motivating power of normative expectations, or others' resentment, is greater than self-resentment. Others' resentment causes embarrassment and shame in the violator. These emotional sanctions work as norm-enforcers, and it is the aversion toward such emotions, rather than some sort of intrinsic motivation, that generally motivates norm compliance. If aversion to others' resentment—as opposed to self-resentment or other intrinsic motivations—has generally more grips on norm compliance, then people will tend to be less norm compliant or behave much less prosocially in anonymous or private conditions compared with what they do publicly.

There is experimental evidence that participants of economic games tend to behave more selfishly or merely to appear to be fair without being fair, when their choices cannot be detected by other players (Bicchieri & Chavez, 2010; Dana, Weber, & Kuang, 2007). Norm abidance and prosocial behavior would thus depend more on what other people expect from the decision maker than on some intrinsic motivation.

There are two problems with this criticism, however. First, a large number of studies in experimental economics also show that people are often motivated to repay gifts and punish violations of certain social norms in anonymous one-shot interactions with genetically unrelated strangers, even at substantial costs to themselves (Fehr, Fischbacher, & Gächter, 2002; Gintis, Bowles, Boyd, & Fehr, 2003). Even in games with asymmetric information like Dana et al.'s (2007), a significant proportion of participants behave prosocially both in public and in private conditions. This body of evidence indicates that, in experimental situations, people generally comply with social norms or behave prosocially, not only because they are averse to others' resentment, but also out of intrinsic motives.

In real-life situations, aversion to others' resentment can have a more motivational grip than self-resentment or other sorts of intrinsic motives, depending on situational cues and the information available (Cialdini & Goldstein, 2004). People often care about complying with social norms independently of what others expect them to do.

Second, conceptually, it seems that people should already care about others' normative expectations for those emotions to have grip on their minds. If I don't care about others' expectations, preferences, and behavior in a certain situation, then I shall be indifferent to their resentment. Along these lines, Anderson concludes "[emotional] sanctions are only a supplementary motive to the original motive for compliance, without which the norm would never have been established" (Anderson, 2000, p. 184).

What I wish to emphasize with Anderson's argument is that, psychologically, the motivational source of compliance appears to reside in the capacity to care. To illustrate and give grounds for this point I now provide some counterexamples to Sugden's resentment hypothesis (1998, 2000). The bottom line is that caring for

other people's preferences, expectations, and behavior is a necessary condition for the arousal of resentment—and for norm compliance, too.

Consider this situation. After their weekly reading group, the participants regularly go to the pub. Ana Maria and Angelica are two of the reading-group members who normally go to the pub. "Going to the pub" and "Not going to the pub" are alternative actions open to Ana Maria in that type of situation. Within the group, it is common knowledge that a person in Ana Maria's position normally goes to the pub rather than not. It is common knowledge that Angelica has good grounds to expect that Ana Maria will go to the pub. It is also common knowledge that people in Angelica's position prefer that people in Ana Maria's position go to the pub rather than not. Would this be sufficient for Angelica to feel resentment if Ana Maria doesn't go to the pub today after their reading group?

It does not seem so. Robert Sugden's (1998, 2000) resentment hypothesis is fulfilled, yet this fails to qualify as a case where resentment is aroused. Ana Maria and Angelica are not close friends; Angelica might be surprised or curious for why Ana Maria is not going to the pub, but she hardly will resent her. To explain why I don't think Angelica would resent Ana Maria, consider another situation.

It's Kirsty's birthday, and Rhiannon is Kirsty's best friend. Kirsty has invited Rhiannon to her birthday party. "Going to the party" and "Not going to the party" are alternative actions open to Rhiannon. Would the resentment hypothesis be sufficient for the arousal of resentment in Kirsty if Rhiannon doesn't go to her party? It is reasonable to believe that in this case, Kirsty would feel resentment. In contrast to the situation above, Kirsty and Rhiannon are friends and they care for each other. To better illustrate the relevant phenomenon of caring, I point to yet another example.

I live in the Edinburgh area. I read in the newspaper that Miss Carr was found driving on the wrong side of the road in Leith, which is part of the Edinburgh area. I don't know anyone in Leith, I have never been there, and don't plan to go there. Is it plausible that I will feel resentment—in Sugden's (2000) sense—toward Miss Carr? Again, I think it is not. In this case, both Miss Carr and I are part of the general population, *P*, of drivers in the Edinburgh area. *P* is quite large. The conditions in Sugden's resentment hypothesis are fulfilled, yet it would be implausible to think that I would feel "a sensation or sentiment which compounds disappointment at the frustration of one's expectations with anger and hostility directed at the person who is frustrating (or has frustrated) them" (Sugden, 2000, p. 113). I won't resent Miss Carr because her behavior does not matter to me, even though she frustrates my expectations, and I may interact with her in the future.

This last example illustrates that, in real-life situations, when we deal with people who are less close to us, we tend to care less about their preferences, expectations, and behaviors. Such people are typically members of other groups, so they are not close to us also in a literal sense: both spatially and temporally (on ingroup–outgroup and social preference, see Bernhard, Fehr, & Fischbacher, 2006; Chen & Xin Li, 2009). In real-life situations, especially when a population is large and it is unlikely that one individual, *i*, will come to know and interact personally with another individual, *j*, *i* will not tend to resent actions by *j* that frustrate her expectations. If it is unlikely that *i* will come to know and interact personally with either *j* or a third individual, *k*, then *i* will not tend to resent actions by *j* that frustrate *k*'s and her expectations.

We care more for people to whom we are close, people we regard as important to ourselves. And the evidence that experimental participants often have an intrinsic motivation to comply with social norms is consistent with this claim. In experimental settings that more closely resemble everyday life, participants tend to behave more generously with closer individuals indeed (Hoffman, McCabe, & Smith, 1996; Charness & Gneezy, 2008).

If the analysis of these cases is correct, then Sugden's hypothesis is insufficient for the arousal of resentment. Feelings of resentment do not simply arise in an individual because her expectations are disappointed. People feel emotions only about things that matter to them, things they care about. If people feel no emotion about things about which they don't care, then they will feel resented when their expectations are disappointed only if they care about the object of those expectations. If feeling resentment and aversion of being the

focus of others' resentments depends on caring, then the resentment hypothesis is not sufficient to explain general norm-abiding behavior.

1.2 Caring and Preferring

Here is an objection to the claim that the resentment hypothesis is insufficient because people feel emotions only about things they care about: "Preference" can be considered a free parameter in Sugden's (2000) account and can take different strengths. Caring about something would amount to having a strong preference for that something, and so P3 would be false—and P4 and P4′ would be incoherent. Therefore my argument would be entirely consistent with Sugden's resentment hypothesis.

This objection is problematic however. To begin with, even if we agree that preferences have different strengths and that "care" can be treated as "strong preference," nothing in Sugden's formulation of the resentment hypothesis suggests how to identify an adequate threshold for the preference parameter. An individual *j* may prefer that people in *i*'s position do Φ rather than Ψ. Still, these pairwise preferences (for *i* doing Φ over *i* doing Ψ) might remain below a certain threshold. For example, if the strength of a preference is measured on an interval from 0 to 1, the preference of *j* for *i* doing Φ can be 0.2 while *j*'s preference for *i* doing Ψ can be 0.1. Sugden's conditions are satisfied, but if the preferences are so weak, it seems implausible to think that the resentment hypothesis is sufficient to raise resentment in *j* when *i* does Ψ instead of Φ. A further condition is required in Sugden's formulation that specifies a suitable threshold, such that *i*'s preferences, expectations, and decisions do matter to *j*.

It may be protested that we could empirically uncover the value of the preference parameter, such that if one individual's preference is unsatisfied, she or he will feel resentment. Different people may care more or less about the expectations and beliefs of others, or perhaps in different situations we care more than in others. By examining the correlations between choice behavior and nonchoice data such as emotional reactions in a given context, a threshold for the preference parameter for resentment arousal could be identified. In this sense, Sugden's (2000) account is sufficient as it stands, and the concepts of caring and preferring would be assumed to be identical.

Another objection to my argument is as follows.[1] Accounts of social behavior such as Sugden's should be interpreted as being restricted to those actions about which one *ought to* have preferences. Such restrictions of domain aren't unusual in economics and philosophy: many economists, for instance, assume that preferences only matter if they affect one's personal welfare, and that the only preferences of this type are self-regarding preferences. In the philosophical literature, social choice functions may rule out preferences over some actions or options, ostensibly on the grounds that such preferences ought not to be considered or ought not to exist (e.g., Sen, 1977; Griffin, 1986; Hausman, 2012). If we think that people shouldn't have preferences over my actions if they don't care about my actions, then this is a small revision to Sugden's resentment hypothesis.

The problem here is that assuming Sugden's account is restricted to preferences or expectations one *ought to* have would defy the very aim of Sugden's account. As pointed out above, the resentment hypothesis is an empirical, descriptive hypothesis, and aims to explain norm compliance with no appeal to normative concepts. Accordingly, the account is not so much concerned with the question of when one should resent, or what types of preferences and expectations one should have to reasonably want to comply with a social norm.

Now, if preferring is not the same as caring—as P3 above claims—then there are grounds to argue that the resentment hypothesis is insufficient. For it might be the case that, as claimed by P4 and P4′, in some sense, *j*'s preference about *i*'s behavior and expectations are strong, but those expectations and behaviors don't really matter to *j*; *j* doesn't care about them. If one does not feel resentment about something unless it matters to him or her, then *j* won't feel resentment when her preferences are frustrated.

There are two questions then: First, what does it mean to care about something? Second, what is the relationship between caring and preferring? Would it make sense to say that an

[1] I am grateful to an anonymous referee for pressing this point.

individual (strongly) prefers A over B and yet she doesn't really care about A? My answers to these questions heavily rely on Harry Frankfurt's (1982, 2004) analysis of caring. Let's start from the latter question.

To care about something is not simply to prefer, desire, or want it. Attributing a preference "to a person does not in itself convey that the person cares about the object" prefers (Frankfurt, 2004, p. 11). Many of our preferences and desires are utterly insignificant. We don't care about them. Satisfying them is of little or no importance to us. For example, I prefer to drink water over coke at this very moment. As I am drinking coke, my preference is unsatisfied. But I don't feel any frustration since I don't really care about such a preference. Note, however, that my drinking coke now does make some difference to me, which suggests that things we deem important to us, and hence things we care about, are not simply things that make some difference to us. Having coke and not water right now is a difference inconsequential to me. As argued by Frankfurt, "nothing is important unless the difference it makes is an important one" (Frankfurt, 1982, p. 259).

This lack of caring and frustration need not be because my preference is weak, or has low intensity. "Sheer intensity . . . implies nothing as to whether we really care about what we want." Frankfurt goes on to explain: "Differences in strengths of desires . . . may be radically incommensurate with the relative importance to us of the desired objects" (Frankfurt, 1982, p. 259). In the case of preference, from the higher strength of my preference for reading a book over doing the laundry, it does not follow that I especially care about the object of this preference. Even if I intensely prefer one over the other, the difference between reading a book instead of doing the laundry is not especially important to me now.

Furthermore, "a person who wants one thing more than another may not regard the former as being any more important to him than the latter" (Frankfurt, 1982, p. 12). Frankfurt makes this claim stick with an example. Suppose that you need to kill time and you decide to watch the TV. You start to watch a certain program because you prefer it to the others that are available. "We cannot legitimately conclude that watching this program is something that [you] care about." After all you are killing time. "The fact that you prefer it to the others does not necessarily entail that you care more about watching it than about watching them, because it does not entail that you care about watching it at all" (Frankfurt, 1982, p. 12). By the same argument, the fact that the individual i prefers that j does Φ rather than Ψ does not entail that i cares more, or at all, about j doing Φ than j doing Ψ.

Suggesting that preferring and caring are distinct concepts is also the empirical finding that our preferences are subject to powerful contextual influences (Lichtenstein & Slovic, 2006). There may not be stable facts about one's preferences independent of the way a given choice situation is framed. Caring—understood in a way to be made clearer in a moment—is relatively stable. "A person can care about something over some more or less extended period of time. It is possible to desire something, or to think it valuable only for a moment . . . But the notion of caring implies a certain consistency or steadiness of behavior; and this presupposes some degree of persistence" (Frankfurt, 1982, p. 261).

If caring and preferring are distinct, what does it mean to care about something? To care about something is not simply to desire it, or want it, or prefer it over something else. Caring is not the same as factoring in things that make some difference. Caring is a mode of the will: "caring about something may be a complex mode of wanting it" (Frankfurt, 2004, p. 11). For Frankfurt, the capacity to care about something should be understood more precisely as the capacity to commit ourselves to our own desires, wants, and preferences. When people care about something they desire to have a desire for it, and they endorse such desire. If a person cares about something, then she is willingly committed to her desire about that thing: she desires that she desires it (Ibid., p. 16). Thus, Frankfurt explains: "by its very nature, caring manifests and depends upon our distinctive capacity to have thoughts, desires, and attitudes that are about our own attitudes, desires, and thoughts" (Frankfurt, 2004, p. 17).

This does not mean that all we fundamentally care about is ourselves and our well-being. It does not implicate that we care about other people's preferences, expectations, and behavior only because we care about our welfare and well-being. I can care about a waitress expect-

ing me to leave a tip after my dinner, and comply with a norm of tipping, even if I am aware that by meeting her expectation I won't feel or be any better off.

In sum, according to Frankfurt, "these alternative possibilities—commitment to one's own desires or an absence of commitment to them—define the difference between caring and not caring" (Frankfurt, 2004, p. 21). It should be clear that, as Frankfurt characterizes it, caring about something is peculiar to humans because it requires the ability to reflexively deal with higher order desires. This ability should depend on reflexive thinking, and on the ability to commit oneself resiliently to distinct courses of action. So strictly speaking, nonhuman animals cannot care in Frankfurt's sense.

Yet, it is plausible to understand caring more broadly than Frankfurt, so that we can make room for the possibility of nonhuman animals that care. Fisher and Tronto (1990) offer a broader characterization, according to which caring is "a species of activity that includes everything we do to maintain, contain, and repair our 'world' so that we can live in it as well as possible. That world includes our bodies, ourselves, and our environments, all of which we seek to interweave in a complex, life-sustaining web" (Fisher & Tronto, 1990, p. 40).

This definition is consistent with Frankfurt's: It construes caring as a complex activity supported and informed by a commitment to those desires and goals that we deem important in life. According to this definition, however, commitment to one's own desires need not be reflexive nor involve self-awareness or consciousness. Agents can hold on to their desires in a persistent steady way without being conscious of their commitments. Caring in this sense corresponds to a relatively stable volitional profile, something through which agents steadily orient in their behavior and their environment in pursuit of a good life.

Also nonhuman animals, in this sense, would have the capacity to care. Nonhuman animals care about staying alive, about avoiding injuries, predators, hunger, thirst, and disorder; they may care about close kin, friends, and other members of their groups; and some may even care for strangers under certain circumstances (cf. Churchland, 2011, Ch. 3). So, caring can be both self- and other-directed. Both humans and some nonhuman animals can care not only about themselves, but also about others.

It is not obvious what mechanism could ground the capacity to care. Patricia Churchland (2011) has suggested that hormones such as oxytocin and vasopressin, which originally evolved to promote self-preservation and care for offspring, probably constitute basic features of the mechanism for caring. In their evolutionary trajectory, these hormones would have later been coopted to serve new jobs to enable wider forms of sociability, and ultimately to foster moral cognition. The last section of this paper integrates Churchland's proposal by pointing to some neurocomputational features of a putative mechanism for caring.

2 Hedonism and Norm Compliance

Sugden's (1998, 2000) argument focuses on negative emotions like anger, fear, and resentment. Even if his argument fails, it is possible that positive emotions are the ultimate motivational source of norm compliance. Pleasure is the main candidate, since it has traditionally been linked to motivation in theoretical as well as empirical accounts of decision making (e.g., Bentham, 1789/1970; Wise, 1982).

Recently, it has been suggested that data on the neurobiological processes underlying social preference are best understood in hedonic terms. From this perspective, pleasure is the ultimate motivation for norm compliance. Theories of social preference model how people rank allocations of material payoff to self and others during strategic interaction (Fehr, 2009). According to these theories, individuals are also concerned with the payoff, preferences, and beliefs of other individuals. Theories of social preferences are not committed to any specific interpretation of the processes underlying decision-making. In particular they do not make any claim with respect to the hedonic significance of norm compliance behavior.

Ernst Fehr and Colin Camerer (2007) have argued that a hedonic interpretation of theories of social preference nicely fits the neurobiological data on social norm compliance. They draw on experimental findings from neuroeconomics to support the claim that individuals derive "higher hedonic value" (p. 420) from outcomes associated to the decision to comply with norms of cooperation or fairness. Fehr and Camerer's

(2007) argument can be reconstructed as follows.

P1. Norm compliance, in general, and "altruistic, fair and trusting behaviors" in particular, "are consistently associated" with neural activity in the striatum (Fehr and Camerer, 2007, p. 419).
P2. Activity in the striatum represents anticipated or experienced reward.
C1. Norm compliance is rewarding.
C2. People comply with social norms because it is rewarding.

If we assume that reward is just hedonic value or pleasure, then C2 is a version of motivational hedonism. Motivational hedonism, in its strongest formulation, is the claim that only pleasure (or pain) motivates us. Fehr and Camerer (2007) do not claim that the evidence they review is sufficient to establish C2. Still, they claim that the evidence strongly supports the hypothesis that norm compliance and prosocial behavior have special reward value. To evaluate Fehr and Camerer's argument, two questions need be answered: First, what does reward amount to? Second, can the same data considered by Fehr and Camerer be more plausibly explained with no appeal to pleasure? After having recalled the types of findings that purportedly support C1, I engage P2 by considering different computational roles of the striatum. Different meanings of "reward" are distinguished, and I argue that pleasure does not motivate norm compliance.

Fehr and Camerer (2007; p. 420) review evidence from three types of sources. First, they cite an unpublished work by Kosfeld, Fehr, and Weibull in which questionnaire data would support the view that "mutual cooperation in social exchanges has special subjective value, beyond the value that is associated with monetary earnings." Second, they survey the findings of a number of neuroimaging experiments in which striatum activity has been observed to be significantly correlated with cooperative outcomes. Third, they notice that striatal activity in one experimental condition can be used to predict choice behavior in a different experimental condition, thereby lending support to C2, that is to the claim that norm compliance occurs because it is rewarding.

Because Fehr and Camerer (2007) do not provide details of Kosfeld et al.'s questionnaire, it is difficult to assess to what extent a hedonic component plays a role in the subjects' ratings. There is some evidence that dopamine activity does not most reliably correlate with ratings of the hedonic experience associated with a drug. For example, in spite of significant loss of most dopamine neurons in the basal ganglia, patients with Parkinson's disease have been reported to have normal subjective pleasure ratings for sweet food (Sienkiewicz-Jarosz et al., 2005).

However, the strongest reason provided by Fehr and Camerer (2007) in support of hedonic interpretations of theories of social preference is that other-regarding and norm-compliance behaviors are consistently associated with activation in the striatum. The striatum is part of what is called the "reward circuit." Camerer and Fehr interpret the processes carried out by activity in this area in terms of hedonic processes. But "reward" and "hedonic processes" are equivocal. In light of current evidence, the computational role of the striatum is complex and may well comprise a number of subcomputational routines. A brief description of the anatomy of the striatum should highlight this last point.

The striatum is a subcortical part of the brain. It is the main input station of the basal ganglia, which are primarily implicated in motor control, learning, and decision-making. Because dopamine is the major striatal neuromodulator, the striatum is thought to be one of the main hubs of the reward circuit. However, it is not the only area associated with reward processing: The ventral tegmental area, the amygdala, the prefrontal cortex, and certain parts of the thalamus are also involved in reward processing (see, e.g., Schultz, 2007a, 2007b). The ventral part of the striatum consists of the caudate nucleus and the putamen. The ventral striatum—or nucleus accumbens—constitutes a third subdivision of the striatum. These three striatal regions are anatomically and functionally distinct. Current evidence suggests that discrete regions of the striatum are differentially involved in the integration of sensorimotor, cognitive, and emotional information, and in action selection and initiation (Knutson, Delgado, & Phillips, 2009). But what does it mean that the striatum processes rewards?

Here are examples of rewards that seem to be processed by such a circuit: sweet tastes, co-

caine, sex, money, smiling faces, and norm compliance. In a general sense, rewards can be understood as objects or states that make us come back for more. In a narrower sense, reward refers to subpersonal informational signals that play specific roles in reinforcement-learning (RL) algorithms, which might be implemented by certain populations of neurons.

RL is a field in computer science and machine learning that offers a collection of algorithms to address the problem of learning what to do in the face of rewards and punishments received by taking different actions in an unfamiliar environment (Sutton & Barto, 1998). A wealth of evidence indicates that activity of dopaminergic neurons in the basal ganglia can be described as implementing a reward-prediction error, which is a signal used by some classes of RL algorithms (Montague, Dayan, & Sejnowski, 1996; Niv, 2009; Glimcher, 2011). A reward-prediction error is the difference between obtained and expected reward. To say that dopaminergic neuronal activity can be described as implementing a reward-prediction error is to say that some neurons can be described as performing computations by executing some RL algorithm. By executing this algorithm, the brain carries out the cognitive task of learning what to do in the face of rewards and punishments obtained in an unfamiliar environment.

Reward as a psychological notion has distinct aspects. The neuroscientist Kent Berridge identifies three dissociable aspects of reward: liking, wanting, and learning (e.g., Berridge, 2003; Berridge, Robinson, & Aldridge, 2009). "Liking" refers to the hedonic experience of a subject. Reward here is a state or outcome that generates a pleasant feeling. "Wanting" (or "incentive salience") refers to a drive toward the pursuit and/or consumption of some typically salient state or outcome. It need not be conscious. Reward in this sense is what is desired, often unconsciously, regardless of its hedonic properties. Learning involves the capacity to associate stimuli and actions to consequences. Reward here consists in states, events, and stimuli that guide agents' learning.

In light of these distinctions, to say that the striatum processes reward can mean at least three different things. Since the relevant sense for Fehr and Camerer's (2007) argument is the hedonic one, we should read P2 above as: Activity in the stratum represents anticipated or experienced pleasure. To assess P2, we should then turn to consider the evidence about the neurobiological underpinnings of hedonic experience.

Berridge and collaborators (Berridge, 2003, 2007; Berridge & Kringelbach, 2008; Berridge, Robinson, & Aldridge, 2009) provide substantial evidence that liking or hedonic experience is generated by opioid, endocannabinoid, and gamma-aminobutyric acid (GABA)–benzodiazepine neurotransmitter systems. Two "hedonic hot spots" have been found in the nucleus accumbens and the ventral pallidum, respectively, which are two regions of the striatum. The first hot spot comprises 10% of the volume of the nucleus accumbens, a relatively small portion of the striatum. Outside those hot spots, in the same two regions, opioids do not enhance liking: Enhancement of hedonic experience is then anatomically restricted to small portions of the striatum (Berridge & Kringelbach, 2008). Hence, the claim that the striatum is a hedonic area need be strongly qualified.

One reason for interpreting the striatum as a pleasure center has been that its major afferents come from dopamine neurons, which have been traditionally considered "pleasure neurotransmitters." However, manipulations of dopamine activity do not appear to have systematic effects on hedonic experience (Berridge & Robinson, 1998). This suggests that dopamine activity is neither necessary nor sufficient for generating hedonic experience, but psychologically, is probably necessary for "wanting," and neurocomputationally, for implementing certain forms of RL algorithms (see Berridge, 2007 on "wanting;" Schultz, 2007b for distinct computational roles of dopamine).

If hedonic experience is the ultimate motive of norm compliance, pleasure should be the triggering cause of the selection of a certain action. Pleasure should come before "wanting" and should determine what we do. Evidence needed to confirm or refute this claim might be gained by focusing on the causal relationship between mechanisms of action selection and "liking."

Computational models of the basal ganglia, of which the striatum is the major nucleus, provide one way to approach this issue. In the framework of RL, one of the best models of the basal ganglia mechanism is the *actor–critic* architecture (Houk, 2007). This class of models

seems to capture some principles of dopaminergically controlled plasticity in the striatum (e.g., Joel, Niv, & Ruppin, 2002). In such models, an actor selects the action to be taken, given the current input, whereas the critic drives the learning process by assessing how well the outcome of one action tallies with the attainment of a certain goal. Neurobiological data suggests that the ventral striatum is associated with the critic, and the dorsal striatum is associated with the actor (Daw, Niv, & Dayan, 2005, Sec. 4). Although these are simplified models, if compared with the complex anatomy and physiology of the striatum, we can still draw some conclusions about the relationship between action selection and pleasure.

From the characterization of the computational architecture of the striatum and the localization of hedonic hot spots thereof, it seems that the main computational business of the striatum is not hedonic processing. Hedonic hot spots might be activated after the critic computes the extent to which the outcome of the action taken matches that which was expected. Theoretically, they register the pleasure of learning rather than the drive of learning. If this is so, then hedonic experience would be the output of decision-making systems over which pleasure has no direct control. The causal interplay between learning and what we want would then make pleasure a contingent result of the appraisal of the outcomes of our actions. Pleasure, therefore, is probably not the ultimate motive of norm compliance: People generally do not comply with norms because it feels good.

3 Neurocomputation and Caring

Caring—if we agree with Frankfurt (1982)—is a complex mode of the will. It corresponds to a relatively stable volitional profile that steadily and persistently orients and guides one's behavior toward self and others in the pursuit of a good, adapted life. I conclude this paper by sketching a possible mechanism for the capacity to care. The proposal has two parts. The first component is anatomical, and identifies some core brain circuits that might support the capacity to care. The second component is physiological, and includes the main neuromodulators along with their computational significance by which caring might be regulated.

The goal is not to provide a complete overview on the anatomy, neurophysiology and computational functions underlying the capacity to care, but rather to elucidate caring by characterizing some of its core mechanistic components. The account is tentative, crude, and incomplete, but it provides the grounds for formulating several empirical questions—some of which will be made explicit in the Conclusion—concerning the relationship between norm compliance, emotion, and caring.

3.1 Anatomy

The amygdala and the ventromedial prefrontal cortex (vmPFC) might be the anatomical core of the capacity to care. Evidence for this hypothesis is given by the fact that lack of caring is one of the prominent traits of a psychopath, in whom the normal functioning of amygdala and vmPFC appears to be disrupted, as well as of patients with ventromedial frontal lobe damage. Psychopathic and ventromedial patients are similar in displaying an abnormal volitional, motivational, and emotional profile. Typically, they make poor choices, have emotional dysfunctions characterized by shallow affect, reduced autonomic response to a variety of stimuli, lack of empathy and attachment to others; they act maladaptively in social situations, behave irresponsibly and impulsively, and present poor behavioral control (Blair, Mitchell, & Blair, 2005; Damasio, 1994).

Psychopathy is a developmental disorder that involves pathological social behavior. Psychopaths routinely violate important norms in society. They are glib, impulsive, irresponsible, manipulative, egocentric, callous, lack empathy, and have shallow emotions (Hare, 2003). Current evidence indicates that they know that their behavior is immoral (Schaich Borg & Sinnott-Armstrong, in press). They do not lack intelligence; what they lack is appropriate volition, motivation, and control: They simply do not care.

Although psychopathy is not a neurological condition, two brain circuits appear to be reliably involved in psychopathic behavior: the amygdala and the vmPFC (Blair, 2007). The amygdala is a complex structure with a wide range of connections with other brain regions, including cortical and subcortical networks. Recent literature has understood the functional roles of the amygdala in a broader and more abstract way than emotional

processing. The type of functional role played by amygdala's processing would be to coding for biological and social relevance (Sander, Grafman, & Zalla, 2003). If the amygdala is a relevance detector, then its processes would have multiple dimensions,

> Including processing of salience, significance, ambiguity, unpredictability and other aspects of 'biological value.' More broadly . . . the amygdala [would have] a key role in solving the following problem: How can a limited capacity information processing system that receives a constant stream of diverse inputs selectively process those inputs that are the most relevant to the goals of the animal? (Pessoa & Adolphs, 2010, p. 780).

If caring for something or somebody involves a resilient commitment to relevant desires and goals important for our lives, then the amygdala would be a core component of the mechanism of caring. The amygdala would contribute to bestowing importance to particular objects, actions, and people, thereby, facilitating the allocation of processing resources to those inputs in a given situation that are the most relevant to the goals and desires of the agent.

The vmPFC is structurally and functionally connected to several brain areas, one of which is the amygdala. Damage to the vmPFC appears to be associated with some symptoms that may be found in psychopaths. Ventromedial patients retain IQ, and perform normally in many standard neuropsychological laboratory tests for reasoning and working memory, but these patients seem less able to feel emotions when faced with some stimuli that reliably elicit emotions in normal people. Like psychopaths, they are capable of making appropriate moral and social judgments. Unlike psychopaths, however, they do not typically exhibit violent behavior.

Antonio Damasio (1996) observes that ventromedial patients

> Have difficulty planning their work day; difficulty planning their future over immediate, medium, and long ranges, and difficulty choosing suitable friends, partners, and activities. The plans they organize, the persons they elect to join, or the activities they undertake often lead to financial losses, losses in social standing, and losses to family and friends. The choices these patients make are no longer personally advantageous, are socially inadequate, and are demonstrably different from the choices the patients were known to have made in the premorbid period (p. 1413).

So, after ventromedial damage, it appears that the capacity to care is severely compromised. Being recruited along with the amygdala in a network that codes for and weighs the value of different biological and social stimuli, the vmPFC might enable selection of appropriate actions on the basis of information relevant to the goals and desires important to the agent's life.

3.2 Physiology

Even if anatomical structures such as the amygdala and vmPFC are intact, a functional capacity for caring requires that the levels of various neuromodulators are maintained within certain bounds. Neuromodulators are neurotransmitters that have spatially distributed and temporally extended effects on their receptors. They affect globally and at longer time scale the computations that brains carry out. Learning, motivation and decision-making are three behavioral functions where neuromodulators are deeply involved (cf. Doya, 2002; Dayan, 2012).

The main neuromodulators by which caring might be regulated include dopamine—which was discussed in Section 2—acetylcholine, serotonin, and noradrenaline. Acetylcholine modulates synaptic plasticity in the cerebral cortex, striatum, amygdala, and hippocampus. Depletion of acetylcholine neurons is associated to memory disorders: it may impair working memory for some stimuli and acquisition of new information (Hasselmo, 2006). More relevant to the capacity to care, acetylcholine may modulate "the information coding in the cortex and the hippocampus so that their response properties are not simply determined by the statistics of the sensory input but are also dependent on the importance of the sensory inputs" (Doya, 2002, p. 503). The idea is that the acetylcholinergic system controls how quickly old information is updated by experience, affecting the storage and update dynamics of memory at both cellular and circuit levels. Since caring is a relatively stable volitional profile, it requires a balance between the retention of old memories and acquisition of new information. If what has already been learned is overwritten too quickly or too slowly, then the pursuit of the goals and desires important for our life will become unstable or inflexible.

The main source of noradrenaline (or norepinephrine) is the locus coeruleus. The action of noradrenaline on its multiple targets varies enormously in different brain regions. Generically, noradrenaline enhances vigilance and

arousal, especially in urgent situations. It facilitates flexible interactions with an uncertain environment (Yu & Dayan, 2005). Blocking noradrenergic inputs to the forebrain may cause deficit in shifting attention between stimuli relevant to carry out a task. Shift of attention to the stimuli more relevant to a given task is facilitated by enhance noradrenergic input (Sara, 2009). If what we care about is positively correlated to its urgency, and caring involves higher vigilance, then acting on what we care about is subject to control by the noradrenergic system.

The neurons of the raphe nuclei are the principal source of serotonin in the brain. The serotonergic system has been characterized as an opponent of the dopaminergic system, involved in computation of aversive outcomes and inhibiting behavior. Low serotonin levels are associated with poor impulse inhibition and negative behavioral reactions; but the opposition between dopaminergic and serotonergic systems is more complex, unsymmetrical, and dynamic (Dayan, 2012). One idea about serotonin's role in caring is that impulsive behavior may be triggered by a decrease in the importance of outcomes distant in the future compared with more proximal ones (Doya, 2002). Higher levels of serotonin would lead agents to be sensitive to reward predictions longer in the future. Low levels of serotonin instead would lead agents to be insensitive to larger delayed rewards and so to behave impulsively. So, insofar as caring involves conflicts and trade-offs between immediate and long-term outcomes, serotonin levels will regulate what we care about at a given time, and how stable our caring about it is over time.

Provided the complicated and dynamic interactions between different neuromodulators, there will be many different combinations of neuromodulators levels that will support a functional capacity for caring. This capacity for caring will be necessary to comply with norms, to enjoy a functional emotional life, and to adaptively navigate our social environment.

Conclusion

This paper has focused on two arguments about the relationships between emotion and norm compliance: Robert Sugden's resentment hypothesis, and Ernst Fehr and Colin Camerer's hedonistic hypothesis. I argued that both arguments are unconvincing. This paper contributes to current literature by explicating the notion of caring, at both the personal and subpersonal level of explanation, suggesting that caring might be the source of both feeling emotions and complying with norms. Questions for further research motivated by the present contribution include: Will agents with impairments in experiencing pleasure or resentment have relatively more difficulty in complying with social norms? How could care be modeled (cf. van Staveren, 2005)? What is the precise conceptual relationship between caring and preferring? What are the more fine-grained psychological dimensions of caring? What anatomical and physiological differences contribute to variation in caring about own and others' welfare? How do these differences impact emotional experience?

References

Adolphs, R. (2010). Emotion. *Current Biology, 20,* R549–R552. doi:10.1016/j.cub.2010.05.046

Anderson, E. (2000). Beyond Homo economicus: New developments in theories of social norms. *Philosophy and Public Affairs, 29,* 170–200. doi:10.1111/j.1088-4963.2000.00170.x

Bennett, M. R., & Hacker, P. M. S. (2003). *Philosophical foundations of neuroscience*. Oxford, UK: Blackwell.

Bentham, J. (1970). *An introduction to the principles of morals and legislation*. Edited by J. Burns & H. L. A. Hart. London, UK: Athlone Press. Original published in 1789.

Bernhard, H., Fehr, E., & Fischbacher, U. (2006). Group affiliation and altruistic norm enforcement. *American Economic Review, 96,* 217–221. doi:10.1257/000282806777212594

Berridge, K. C. (2003). Pleasures of the brain. *Brain and Cognition, 52,* 106–128. doi:10.1016/S0278-2626(03)00014-9

Berridge, K. C. (2007). The debate over dopamine's role in reward: The case for incentive salience. *Psychopharmacology, 191,* 391–431. doi:10.1007/s00213-006-0578-x

Berridge, K. C., & Kringelbach, M. L. (2008). Affective neuroscience of pleasure: Reward in humans and animals. *Psychopharmacology, 199,* 457–480. doi:10.1007/s00213-008-1099-6

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research Reviews, 28,* 309–369.

Berridge, K. C., Robinson, T. E., & Aldridge, J. W. (2009). Dissecting components of reward: "liking," "wanting," and learning. *Current Opinion in Pharmacology, 9,* 65–73. doi:10.1016/j.coph.2008.12.014

Bicchieri, C., & Chavez, A. (2010). Behaving as Expected: Public Information and Fairness Norms. *Journal of Behavioral Decision Making, 23,* 161–178. doi:10.1002/bdm.648

Blair, R. J. R. (2007). The amygdala and ventromedial prefrontal cortex in morality and psychopathy. *Trends in Cognitive Sciences, 11,* 387–392. doi:10.1016/j.tics.2007.07.003

Blair, R. J. R., Mitchell, D., & Blair, K. (2005). *The psychopath: Emotion and the Brain*. Malden, MA: Blackwell.

Charness, G., & Gneezy, U. (2008). What's in a name? Anonymity and social distance in dictator and ultimatum games. *Journal of Economic Behavior and Organization, 68,* 29–35. doi:10.1016/j.jebo.2008.03.001

Chen, Y., & Xin Li, S. (2009). Group Identity and Social Preferences. *American Economic Review, 99,* 431–457. doi:10.1257/aer.99.1.431

Churchland, P. S. (2011). *Braintrust: What Neuroscience Tells us about Morality*. Princeton, NJ: Princeton University Press.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology, 55,* 591–621. doi:10.1146/annurev.psych.55.090902.142015

Damasio, A. R. (1994). *Descartes' error*. New York, NY: Putnam.

Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London: Series B. Biological Sciences, 351,* 1413–1420. doi:10.1098/rstb.1996.0125

Dana, J., Weber, R. A., & Kuang, X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory, 33,* 67–80. doi:10.1007/s00199-006-0153-z

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8,* 1704–1711. doi:10.1038/nn1560

Dayan, P. (2012). Twenty-five lessons from computational neuromodulation. *Neuron, 76,* 240–256. doi:10.1016/j.neuron.2012.09.027

de Sousa, R. (2010). Emotion. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring, 2010 ed.). Retrieved from http://plato.stanford.edu/archives/spr2010/entries/emotion/

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks, 15,* 495–506. doi:10.1016/S0893-6080(02)00044-8

Fehr, E. (2009). Social preferences and the brain. In P. W. Glimcher, C. Camerer, R. A. Poldrack, & E. Fehr (Eds.), *Neuroeconomics: Decision making and the brain* (pp. 215–232). New York, NY: Elsevier Academic Press.

Fehr, E., & Camerer, C. (2007). Social neuroeconomics: The neural circuitry of social preferences. *Trends in Cognitive Sciences, 11,* 419–427. doi:10.1016/j.tics.2007.09.002

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation and the enforcement of social norms. *Human Nature, 13,* 1–25. doi:10.1007/s12110-002-1012-7

Fisher, B., & Tronto, J. (1990). Toward a feminist theory of caring. In E. Abel, & M. Nelson (Eds.), *Circles of care* (pp. 36–54.). Albany, NY: SUNY Press.

Frankfurt, H. G. (1982). The importance of what we care about. *Synthese, 53,* 257–272. doi:10.1007/BF00484902

Frankfurt, H. G. (2004). *The reasons of love*. Princeton, NJ: Princeton University Press.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior, 24,* 153–172. doi:10.1016/S1090-5138(02)00157-5

Glimcher, P. W. (2011). Understanding dopamine and RL: The dopamine reward prediction error hypothesis. *PNAS: Proceeding of the National Academy of Sciences of the United States of America, 108,* 15647–15654. doi:10.1073/pnas.1014269108

Griffin, J. (1986). *Well-being: Its meaning, measurement, and moral importance,* Oxford, UK: Clarendon Press.

Hare, R. D. (2003). *The Psychopathy Checklist—Revised,* 2nd Edition. Toronto: Multi-Health Systems.

Hasselmo, M. E. (2006). The role of acetylcholine in learning and memory. *Current Opinion in Neurobiology, 16,* 710–715. doi:10.1016/j.conb.2006.09.002

Hausman, D. (2012). *Preference, value, choice, and welfare*. Cambridge, MA: Cambridge University Press.

Hoffman, E., McCabe, K., & Smith, V. L. (1996). Social distance and other-regarding behavior in dictator games. *American Economic Review, 86,* 653–660.

Houk, J. C. (2007). Models of basal ganglia. *Scholarpedia, 2,* 1633. doi:10.4249/scholarpedia.1633

Huebner, B., Dwyer, S., & Hauser, M. (2009). The role of emotion in moral psychology. *Trends in Cognitive Sciences, 13,* 1–6. doi:10.1016/j.tics.2008.09.006

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor-critic models of the basal ganglia: New anatomical and

computational perspectives. *Neural Networks, 15,* 535–547. doi:10.1016/S0893-6080(02)00047-3

Knutson, B., Delgado, M. R., & Phillips, P. E. M. (2009). Representation of subjective value in the striatum. In P. W. Glimcher, C. Camerer, R. A. Poldrack, & E. Fehr (Eds.), *Neuroeconomics: Decision making and the brain* (pp. 389–406). New York, NY: Elsevier Academic Press. doi:10.1016/B978-0-12-374176-9.00025-7

Lichtenstein, S., & Slovic, P. (Eds.). (2006). *The construction of preference*. New York, NY: Cambridge University Press. doi:10.1017/CBO9780511618031

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *The Journal of Neuroscience, 16,* 1936–1947.

Montague, P. R., & Lohrenz, T. (2007). To detect and correct: Norm violations and their enforcement. *Neuron, 56,* 14–18. doi:10.1016/j.neuron.2007.09.020

Niv, Y. (2009). RL in the brain. *Journal of Mathematical Psychology, 53,* 139–154. doi:10.1016/j.jmp.2008.12.005

Pessoa, L., & Adolphs, R. (2010). Emotion processing and the amygdala: From a "low road" to "many roads" of evaluating biological significance. *Nature Reviews Neuroscience, 11,* 773–783. doi:10.1038/nrn2920

Sander, D., Grafman, J., & Zalla, T. (2003). The human amygdala: An evolved system for relevance detection. *Reviews in the Neurosciences, 14,* 303–316. doi:10.1515/REVNEURO.2003.14.4.303

Sara, S. J. (2009). The locus coeruleus and noradrenergic function, *Nature Reviews in Neuroscience, 10,* 211–223. doi:10.1038/nrn2573

Schaich Borg, J., & Sinnott-Armstrong, W. (in press). Do psychopaths make moral judgments? In K. A. Kiehl, & W. P. Sinnott-Armstrong (Eds.), *The Oxford handbook of psychopathy and law*. New York, NY: Oxford University Press.

Schultz, W. (2007a). Reward. *Scholarpedia, 2,* 1652. doi:10.4249/scholarpedia.1652

Schultz, W. (2007b). Reward signals. *Scholarpedia, 2,* 2184. doi:10.4249/scholarpedia.2184

Sen, A. K. (1977). Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy & Public Affairs, 6,* 317–344.

Sienkiewicz-Jarosz, H., Scinska, A., Kuran, W., Ryglewicz, D., Rogowski, A., Wrobel, E., . . . Bienkowski, P. (2005). Taste responses in patients with Parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry, 76,* 40–46. doi:10.1136/jnnp.2003.033373

Sinnott-Armstrong, W. P. (Ed.). (2008a). Moral psychology*: Vol. 2. The cognitive science of morality: Intuition and diversity*. Cambridge, MA: MIT Press.

Sinnott-Armstrong, W. P. (Ed.). (2008b). Moral psychology*: Vol. 3. The neuroscience of morality: Emotion, brain disorders, and development*. Cambridge, MA: MIT Press.

Smith, A. (1976). *The theory of moral sentiments*. Oxford, UK: Clarendon Press. Original published in 1759.

Sripada, C., & Stich, S. (2007). A framework for the psychology of norms. In P. Carruthers, S. Laurence, and S. Stich (Eds.), *The innate mind: Culture and cognition*. Oxford: Oxford University Press, 280–301. doi:10.1093/acprof:oso/9780195310139.003.0017

Sugden, R. (1998). Normative expectations: The simultaneous evolution of institutions and norms. In A. Ben-Ner, & L. Putterman (Eds.), *Economics, values, and organization* (pp. 73–100). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139174855.004

Sugden, R. (2000). The motivating power of expectations. In J. Nida-Rümelin, & W. Spohn (Eds.), *Rationality, rules and structure* (pp. 103–29). Dordrecht, the Netherlands: Kluwer. doi:10.1007/978-94-015-9616-9_7

Sugden, R. (2002). Beyond sympathy and empathy: Adam Smith's concept of fellow-feeling. *Economics and Philosophy, 18,* 63–87.

Sutton, R. S., & Barto, A. G. (1998). *RL: An introduction*. Cambridge, MA: MIT Press.

van Staveren, I. P. (2005). Modelling care. *Review of Social Economy, 63,* 567–586. doi:10.1080/00346760500364429

Wise, R. (1982). Neuroleptics and operant behavior: The anhedonia hypothesis. *Behavioral and Brain Sciences, 5,* 39–87. doi:10.1017/S0140525X00010372

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron, 46,* 681–692. doi:10.1016/j.neuron.2005.04.026

Received December 31, 2012
Revision received June 6, 2013
Accepted August 22, 2013 ■