Feeling Our Way to the Common Good: Utilitarianism and the Moral Sentiments

The Monist

An International Journal of General Philosophical Inquiry

January, 2010 Volume 93, Number 1

General Topic:

The Meaning of Life

Contributors

Duncan Pritchard
David Heyd & Franklin G. Miller
Lisa Bortolotti
Laurence James
Bence Nanay
Michael Almeida
Tim Oakley
Shidan Lotfi
Jason Burke Murphy
Christoph Fehige & Robert H. Frank

THE MONIST

An International Quarterly Journal of General Philosophical Inquiry

FOUNDED 1888 BY EDWARD C. HEGELER

Editor: BARRY SMITH
Managing Editor: GEORGE A. REISCH
Production: CRAIG W. O'DELL

Historical Note: Paul Carus

Paul Carus, the first editor of *The Monist*, was born in Ilsenberg am Harz on July 18, 1852, and died in La Salle, Illinois on February 11, 1919. After receiving his Ph.D. degree in philosophy and classical philology from Tübingen University in 1876, he taught briefly at the State Military Academy at Dresden. In search of freedom for expression of his independent views, he migrated first to England and then to the United States. In 1887, he accepted the invitation of Edward C. Hegeler (who later became his father-in-law) to edit *The Open Court* magazine, a monthly journal devoted primarily to comparative religion. In 1888, *The Monist* was established as a quarterly journal of the philosophy of science, and Paul Carus served as editor of both journals, and as editor of the Open Court Publishing Company until his death in 1919.

SUBSCRIPTION RATES: United States: Annual (4 issues): Institutions, $55.00; individuals, $35.00; 2 years institutions, $100.00; individuals, $60.00. Single copies: $15.00. Foreign postage: Add $5.00 to single copy rate or $16.00 per year of subscription.

Checks should be made payable to THE MONIST and addressed to THE MONIST, 315 Fifth St., Peru, Illinois 61354

Correspondence concerning manuscripts should be addressed to
Barry Smith, Editor, THE MONIST
Department of Philosophy
University at Buffalo
State University of New York
135 Park Hall
Buffalo, NY 14260–4150 USA

THE MONIST, Vol. 93, No. 1

Peru, Illinois 61354
Published by The Hegeler Institute

VOL. 93, NO. 1 JANUARY 2010

THE MONIST

An International Quarterly Journal of General Philosophical Inquiry

FOUNDED 1888 BY EDWARD C. HEGELER

EDITOR: *Barry Smith*

ADVISORY EDITOR: *Quentin Smith*

MANAGING EDITOR: *George A. Reisch*

GENERAL TOPIC: *The Meaning of Life*

ARTICLES:

Feeling Our Way to the Common Good: Utilitarianism and the Moral Sentiments

The will to conduct one's life with decency has a lot of potential. It could effect peace and coherence in society, and it could effect happiness and satisfaction in those individuals who cultivate it and act on it. For many people, leading a moral life is so important that their will to do so overlaps, or even coincides, with their will to lead a meaningful life. One classical manifestation of the affinity between morality and meaning is the desire to be able to look back one day, from one's death-bed, on a moral achievement that spans one's entire life: on 'leaving the world a better place', for example, or on having done, by and large, the right thing.

Still, on many days the will to morality seems hard to pursue and hard to preserve. Some of the difficulties stem from the fact that the nature of a decent life is unclear. The standards of the good and the right are controversial in the public arena, and no less controversial among the legions of moral philosophers who get paid for elucidating them. The fray over morality both mirrors and constitutes a risk that moral motivation misfires or erodes.

This tract is designed to play a part in reducing the uncertainty and controversy, or at least the risk that they hamper the struggle for a life of value and meaning. We will look at two prominent characterizations of morality: the claim that it is good and right to do what our moral sentiments tell us to do; and the claim that it is good and right to do what would maximize the amount of happiness in the universe. In large parts of the literature the two doctrines are treated as rivals, and a great deal of attention is paid to clashes between them. We will argue that the standard emphasis on clashes is, to say the least, misleading. For the person who tries to lead a moral life the need to choose between the two conceptions of morality is not that urgent. The relationship between the two doctrines is far more symbiotic than it is often held to be. It is our ambition to outline most of that symbiosis, including some parts that other thinkers—as we will point out in the process—have covered before us.

"Feeling Our Way to the Common Good:
Utilitarianism and the Moral Sentiments" by Christoph Fehige & Robert H. Frank,
The Monist, vol. 93, no. 1, pp. 141–165.

1. Clashes

One of the clashes between moral sentiments and the doctrine of utilitarianism that have received so much attention is this. You stand on a bridge over a railroad track and see a trolley-car approach that is out of control. The car is about to strike a group of five people further down the track and would no doubt kill every one of them. You stand behind a fat man whom you could push from the bridge onto the track—an action that would be certain to kill the man, stop the trolley, and save the five other people. There is no other way for you to save them, not even jumping onto the track yourself: you are not fat enough to stop the trolley. Should you push the fat man?[1]

Many people feel that they should not. Utilitarianism, however, says that they should. For the five deaths, we may assume, would involve a larger loss of happiness than the one death, and maximizing happiness is what utilitarianism is all about.

The trolley case is one of several known for having people recoil from the utilitarian verdict.[2] Many of the other cases, too, involve opportunities to save more rather than fewer lives, but to save them in a way that feels objectionable. There is the doctor who could save two patients by painlessly and secretly killing one healthy vagrant and transplanting that person's organs. There is the passenger who after a plane crash could save instead of her own child a distinguished surgeon, who she knows would save many others in turn. Corpulence reappears as a risk factor in the story of the potholers: they could escape drowning by using a stick of dynamite to blast away a fat man who got stuck in the mouth of the cave, trapping them inside while the water is rising. A sheriff can prevent a riot that would involve several deaths, if he gives in to a mob and hangs an innocent man. A botanist in the jungle can, by shooting dead one innocent person himself, prevent a local potentate from having ten innocent people shot dead. And so forth.

Such cases, over which moral sentiment or moral intuition clash or seem to clash with utilitarianism, deserve our attention. But so do the other, happier aspects of the diplomatic relations between the two realms. Those happier aspects, neglected in large parts of the literature, are our topic.

2. Sentiments and Intuitions

Does it matter whether the discussion is conducted in terms of moral sentiments or of moral intuitions? One common way of drawing the dis-

tinction conceives of a moral sentiment as a certain kind of affect that has at least some motivational force: a person feels good or bad about something in a certain kind of way, and this comprises or causes a tendency of hers to bring about or continue that thing in the positive case, or to prevent or stop it in the negative case. In both these respects—feeling and motivation—an intuition can be more detached. A moral intuition is some kind of perception or belief, the content being, for instance, that something is good, bad, right, wrong, just, or unjust. Conceptually speaking, it need not involve either affect or motivation.

Since various parts of this paper will assume that at least one of the two, affect or motivation, are there when the mental states under discussion are, we opt for sentiments rather than intuitions. Whether this conceptual precaution makes much of an empirical difference is another question. Do you know a person who has the intuition that the existence of hunger is bad but lacks either a disposition to a negative affective response to hunger or a disposition to feed the hungry? It seems that moral intuitions rarely come without moral sentiments.

For similar kinds of reasons, we will permit ourselves to use terms like 'moral feeling', 'moral emotion', and 'moral sentiment' by and large synonymously. We acknowledge that a full-fledged moral psychology or philosophy of mind will want to make finer distinctions in this domain, but we conjecture that most of the points in this paper cut, or could be made to cut, across these distinctions.

We will call a moral sentiment utilitarian if it approves of an increase in welfare, or of actions, arrangements, intentions in as far as they point towards such an increase; and non-utilitarian if it doesn't. Non-utilitarian moral sentiments can be concerned with something else altogether (say, with truthfulness) or even be anti-utilitarian in that they approve, at least for a certain context, of a decrease in the overall amount of welfare. The feeling that pushing people from bridges is wrong even if it is the only way to maximize welfare is thus an anti-utilitarian feeling.

3. Lowering the Moral Costs of Decision-Making

Now to the clashes reported in section 1 and to our claim that things are not as bad as they seem. Our first conciliatory remark we borrow from Henry Sidgwick.[3] It has to do less with the axiological input to decisions

than with the most efficient way of reaching them. In emphasizing the costs of decision-making, Sidgwick anticipates work on bounded rationality, satisficing, and the importance of heuristics;[4] in relying partly on the emotions to lessen these costs, he anticipates work in neurobiology and cognitive science.[5] Sidgwick's approach has been fertile in the utilitarian tradition itself as well, where R.M. Hare's theory of two levels of moral thinking is one prominent offspring.[6]

Sidgwick points out that what suggests itself as the utilitarian method can fail to meet a utilitarian criterion. Imagine trying to figure out for a large number of people who of them would with which probability come to enjoy which amount of happiness if you performed possible action 1, possible action 2, and so forth. The deliberative enterprise would often defeat itself, first and foremost by costing attention and money and time that thus become unavailable for the substantial utilitarian task at hand—say, for saving people from starving or drowning. But it would also defeat itself, in spite of all those investments, by breeding errors. The human mind is prone to special pleading and to beliefs that come in handy. For example, an inconvenience to oneself has the remarkable tendency to appear larger than one to the neighbours, and people's ability to correct such appearances for perspective is underdeveloped, especially in times of conflict.

Given the costs to general happiness of the official deliberative route to action, utilitarians can recommend that we employ shortcuts. These shortcuts should be utilitarian in that using them is expected to lead to the largest sum of utility, but non-utilitarian in that they do not appeal to any such sum—for, if they did, they would direct us back on the long way and cease to be shortcuts. By their very nature such shortcuts come with a cost of their own. Since the shortcuts mustn't appeal to sums of utility, cases can come up in which they do not favour the action that would maximize utility. Thus, the utilitarian justification of a shortcut in decision-making will never lie in the inconceivability of such cases, and hardly ever in the nonexistence of such cases in real life, but only in the fact that in real life generally taking the shortcut is likely to produce more utility than generally taking another shortcut or generally taking the long way.

Moral sentiments can be such shortcuts. If they are utilitarian in the one sense we have outlined and non-utilitarian in the other sense we have outlined, the utilitarian will welcome them. It is easy to see how entrusting parts of one's decision-making to the strong feeling that it is wrong to kill

innocent people (for instance, to push them from bridges) has a large expected utility; and similarly for a strong feeling that it is wrong to tell lies or that it is wrong to hire people on grounds other than their competence. In order to fulfil their functions as shortcuts, these feelings must be non-utilitarian to the point of being, one way or the other, anti-utilitarian. They do their job of absolving you from entering into computational mode precisely by leaving no room for utility-based exceptions. They thus have, and must have, a rigidity that amounts to an anti-utilitarian 'even if' or 'no matter whether': it is wrong to push people from bridges no matter whether doing so would maximize happiness.

To be sure, a person who puts such feelings in the driver's seat is bound to get *some* decisions wrong by utilitarian lights, and the decision not to push somebody from a bridge even in the rare situation in which this would save five others is a case in point. But our general remark about exceptions applies. The utilitarian reason for using shortcuts is not invalidated by the fact that shortcuts, qua being short, cannot do justice to every case—especially not if the cases they fail to treat justly are out of the ordinary.

There is another feature of such shortcuts that Sidgwick sees very clearly. They can increase happiness even if the agents who use them are not, deep down, utilitarians and do not use them for utilitarian reasons. In the axiology of those agents themselves, the non-utilitarian sentiments may well have the last word. A person's aversion to telling lies, even if in no way fuelled by a concern for the common good, can still be conducive to the common good. Having considered numerous connections of this kind, Sidgwick concludes that we may "regard the morality of Common Sense as a machinery of rules, habits, and sentiments, roughly and generally but not precisely or completely adapted to the production of the greatest possible happiness for sentient beings". By and large, he says, the machinery is "unconsciously Utilitarian".[7]

Earlier thinkers, too, notice that in many respects utility will be served indirectly, by sentiments or intuitions that have a content other than utility. Adam Smith provides arguments to that effect, although he does not speak as a champion of utilitarianism eager to convince people with non-utilitarian moral sentiments that they need not fear that doctrine. He is a reconciler travelling in the other direction, trying to show to utilitarians that they need not fear his project of using non-utilitarian moral sentiments as the basic and authoritative building blocks of morality.[8] Back in the utilitarian

camp we meet John Austin, the nineteenth-century theorist of jurisprudence, who fully acknowledges that utilitarians would risk thwarting their purpose if they pursued it too single-mindedly. "It was never contended or conceited by a sound, orthodox utilitarian", he writes, "that the lover should kiss his mistress with an eye to the common weal." He discusses the objection that "the occasion for acting *usefully* would slip through our fingers, whilst we weighed, with anxious scrupulosity, the merits of the act and the forbearance". As a remedy against this and other problems, rules and the sentiments associated with them are strongly recommended:

> To think that the theory of utility would *substitute* calculation for sentiment, is a gross and flagrant error: the error of a shallow, precipitate understanding. He who *opposes* calculation and sentiment, opposes the rudder to the sail, or to the breeze which swells the sail. [. . .] To crush the moral sentiments, is not the scope or purpose of the true theory of utility. It seeks to impress those sentiments with a just or beneficent direction: to free us from *groundless* likings, and from the tyranny of senseless antipathies; to fix our love upon the useful, our hate upon the pernicious.[9]

Several decades later, John Stuart Mill points out that people's beliefs about "the effect of things upon their happiness" have a great influence on their sentiments anyway, and thus "a large share in *forming* the moral doctrines even of those who most scornfully reject" the principle of utility.[10] Mill also denies the allegation that the utilitarian doctrine requires its followers to constantly keep their eyes on the "ultimate destination", on "the end and aim of morality". When they "go out upon the sea of life", utilitarians should use "subordinate principles".[11]

Sidgwick, however, discusses indirectness more extensively and with more discernment than any thinker before him, and so we will continue to refer to this kind of approach as his. Indirectness takes some of the sting out of the opposition, or alleged opposition, reported in section 1. What looked like adversity begins to look like division of moral labour. Utilitarianism could *characterize* value and obligation, while the non-utilitarian moral sentiments help *implementing* them.

4. Co-constituting the Good

But how much of a reconciliation can the division of moral labour achieve? Imagine a scheme in which a money-loving slave owner earns an extra dollar every time her slaves harbour thoughts, which they fre-

quently do, that are hostile to her and her wealth. We would hesitate to call that arrangement harmonious. The purely functional fit between utilitarianism on the one hand and non-utilitarian and even anti-utilitarian moral sentiments on the other falls short of our ideals of harmony in just the same way. To be sure, the sentiments help executing the utilitarian programme, but they do so, metaphorically speaking, against their will. Their effects are one thing, their message another. The disharmony that remains is that they speak out against some utilitarian judgments and, to that extent, against utilitarianism.

Dieter Birnbacher and Bernward Gesang argue for a utilitarian move that begins to address that concern.[12] They point out that since people's moral sensibility has a serious impact on their welfare it cannot fail to have a serious impact on the maximizing of their welfare. For example, the fact that feeling moral repulsion lowers a person's welfare makes it unlikely for an action that large parts of the public find repulsive to maximize the general welfare, and thus unlikely for it to be recommended on utilitarian grounds. Utilitarianism, it turns out, has had sensitivity to people's sensitivities, utilitarian or not, built in all along.

The workings of that sensitivity depend on the kind of utilitarianism. A utilitarian formula that takes as its moral currency pleasure and the absence of pain would be sensitive to the pleasure induced in people by the belief that things are run in the way they feel they should be run, and to the pain induced by the opposite belief. Whereas a formula focusing on desire fulfilment may well count the moral sentiments as desires (more on this in section 6) and assess actions that accord or fail to accord with them as actions that raise or lower the general welfare by fulfilling or frustrating the public's desires.[13]

The scope, too, of the utilitarian sensitivity will vary with the moral currency. Consider the example of a repulsive act that would be performed in secret so that there wouldn't be a great deal of felt repulsion, and thus not a great deal of repulsion-induced pain, for a hedonistic utilitarianism to register. In such a utilitarianism, the repulsive act, if it maximizes pleasure in all other respects, would carry the day. However, people's disposition to repulsion (the fact that they would be repulsed if they heard of the act) might amount to a strong implicit desire that the act not happen. If so, even the secret act would frustrate a strong desire harboured by many people, since desires can be frustrated without the desirers ever finding out. In this way, a desire-fulfilment utilitarianism can

give far more weight to the sentiments that speak against the secret act than a hedonistic utilitarianism.

5. Defining and Identifying the Good

Non-utilitarian and even anti-utilitarian moral sentiments, we have just seen, will find their way into the sum of welfare and will thereby deflect the utilitarian verdict in their own direction. Acknowledging this influence is progress not only because it unmasks some alleged clashes as pseudo, but also because it is a first step from exploitation—remember the purely functional move we began with in section 3—to reconciliation proper. With this step utilitarians go beyond treating the moral sentiments merely as a means and begin to give them a say.

Still, if reconciliation is the goal, further steps would be welcome. The moral sentiments have been granted a say *in* the sum of welfare, but still no say *about* the claim that maximizing that sum is the alpha and omega.[14] Clashes remain when moral sentiments have a content that transcends their own impact on happiness—when they have a content of the form: this action is wrong even if, after the happiness-decreasing force of the widespread moral sentiment against it has been taken into account, it remains the action that maximizes happiness.

Is there hope to alleviate that deepest of disharmonies, too? Utilitarians can offer essentially two strategies, which are radical in radically different ways. One strategy is to show that, when it comes to the ultimate characterizing of the good and the right, we are justified in turning a deaf ear to the sentiments. This is R.M. Hare's strategy.[15] In order to discuss its prospects, we would have to ask whether utilitarianism can be given a foundation that gets by without the sentiments—a question too large for this paper.

But here is the opposite strategy. In a sentimentalist metaethics, the sentiments are employed to *define* the good and the right. This can happen in various ways, and we will concentrate on an example of a subjectivist proposal about the definitional link: by calling something good or right a person means to say (roughly) that, if she carefully considered that thing in a cool hour, she would have, all in all, positive feelings (or positive feelings of a certain kind) about it.[16]

Notice that nothing in the sentimentalist proposal as such excludes utilitarianism. A person who looks carefully at various ways the world could be might find that each time, no matter what else is going on, her feelings

end up favouring the scenarios that involve—and because they involve—the largest amount of happiness. A sentimentalist metaethics might well yield utilitarianism as a normative ethics.

There are reasons to suspect more than a logical possibility here. A host of such reasons can be found in the writings of the pioneers of moral sentimentalism themselves, who all have a strongly utilitarian streak. David Hume, for instance, argues in considerable detail for the claim that

> usefulness has, in general, the strongest energy, and most entire command over our sentiments. It must, therefore, be the source of a considerable part of the merit ascribed to humanity, benevolence, friendship, public spirit, and other social virtues of that stamp; as it is the *sole* source of the moral approbation paid to fidelity, justice, veracity, integrity, and those other estimable and useful qualities and principles.[17]

Years before Hume begins to write, Francis Hutcheson already reaches a more radical conclusion. According to Hutcheson, the sentimental data establish that virtue is nothing but "universal benevolence toward all men", a benevolence that aims at "the greatest happiness for the greatest numbers". Hutcheson concedes that in such matters "men must consult their own breast", but argues that if they do they will all find the same.[18]

If that sounds like a bold claim, let us try a weaker one. Consulting our own breast would seem to lead most of us at least into a *ceteris-paribus* utilitarianism. *In as far as* one world contains more happiness than the other, one action leads to more happiness than another, one person tends to do more often than another what she believes would bring about more happiness —*in as far as* these things are the case, our feelings are likely to favour the first world, the first action, the first person.

The major stumbling blocks, then, seem to lie not on the way to *ceteris-paribus* utilitarianism but on the way from '*ceteris paribus*' to 'all things told'. It is here that the *cetera* can fail to be *paria* and that other aspects can ignite sentiments that point in the other direction. We feel that people should not tell lies, should not kill an innocent person, and should give priority to their friends and children or to those who are worst off. These feelings could be strong enough to trump the feelings that favour the greater amount of happiness.

Granted, they could be—but are they in your case? Remember the subjectivism of the metaethical proposal under discussion: you are supposed, when determining *your* normative ethics, to consult *your* feelings. Statistics

about other people's feelings are of no concern in such a project. More precisely, such statistics may form part of the information put before the moral jury (see section 4), but not of the jury itself, which consists of your feelings alone. And those might well lead you into all-things-told utilitarianism.

The more so as the feelings in question must meet the requirement of careful consideration that appears in the sketch of the sentimentalist proposal. The feelings we are dealing with are *about* one thing or another. A feeling is only a feeling about something if it would survive a full and correct and vivid representation of that thing. This is a powerful constraint, especially when the objects are compound or complicated. The more compound or complicated an object, the more danger of misrepresenting or overlooking some of its features and of thus getting an emotional response that is not really a response to that object but to a mutated, or mutilated, version of it.

To get your response to the trolley situation, for example, you need to fully, correctly, and vividly represent to yourself not just the loss of happiness the fat man and his relatives and friends will suffer if you push him—but also the loss of happiness that person number one on the tracks and her relatives and friends will suffer if you don't push the fat man; and that of person number two on the tracks and her relatives and friends; and so forth. An emotional response is a response to the situation (rather than to a distorted or incomplete or otherwise deficient representation of the situation) only if it is, or incorporates, a response to every one of these components. We may well wonder, therefore, how non-utilitarian a response can become before failing to qualify as a response to the entire situation—and to nothing but the situation.[19]

These thoughts on the concept and nature of the good and the right contribute to the reconciliation in another way, too. Belief and evidence sail close behind. If the good and the right *are* a matter of feelings, a promising way to *find out* about the good and the right is to consult the feelings. Epistemological support from the sentiments for utilitarianism would be part and parcel of the semantic or ontological support we have contemplated.

The long and short of this section is that the step from a sentimentalist metaethics to a utilitarian morality could happen in principle, and that it has been taken by some. The feelings of many of us go at least in the same direction, and our belief that they don't go all the way may well be mistaken.

A tension might be thought to loom between this foundational story and the functional story, inspired by Sidgwick, that we set out with in section 3. The earlier story suggests that the feelings help by being non-

utilitarian or even anti-utilitarian, and the current story that they help by being utilitarian. Can these two kinds of reconciliation be reconciled? There is no principled reason why not, provided that both classes of moral sentiments exist and each class plays its own role.

As long as they answer to different standards of representation, the two kinds of sentiments can even coexist within one and the same agent.[20] A sentiment that is supposed to justify a final and thorough moral judgment about a situation must, for that reason, meet high standards of representation, and we have suggested that, in as far as the *controversial* parts of utilitarianism will be endorsed by the sentiments, this will be due largely to that requirement to fully and correctly and vividly take in the situation, and nothing but the situation. Whereas a sentiment that a utilitarian agent deliberately cultivates as a decision-theoretic shortcut must, for that reason, make do with less stringent forms of representation. Digging in its heels after registering just one or two salient features is exactly what it is supposed to do. It must be quick and dirty.

The distinction can be illustrated by the neurobiology of particularly dramatic cases of speed and dirt in the emotions. Joseph LeDoux and others have shown not just that often very little of the information-processing that precedes an emotion is conscious, but that often very little information-processing, conscious or unconscious, precedes the emotion at all. A causal chain from a visual stimulus to a first emotional reaction can simply bypass the department in charge of turning the stimulus into a detailed and accurate representation. This bypassing of the visual cortex allows, if we use LeDoux's favourite example, a hiker's fear of snakes to be triggered even before she believes that she is seeing a snake.[21] There is a point in having the fear set in before the best available representation, just as there is a point in shedding the fear when the best available representation shows the curved object on the ground to be a twig and not a snake. We are not suggesting that in moral dilemmas the time pressure and the neural mechanisms are the same as in encounters with potential snakes. But the point in having two classes of responses, dirty and clean, is of the same kind.

6. Desiring and Having Reasons to Bring about the Good

The previous section was about identifying the good—but what about desiring the good and having reasons to bring it about? It turns out that those two might already be included.

There is a long and strong tradition that sees desires as affects or as tendencies to feel.[22] The core idea is that a person desires something if the thought of it would, literally, please her. The person who would be delighted with the thought that her children will be happy, and sad at the thought that they won't, is the person to whom it is important that her children are happy. Similarly, the child who pictures herself on a new bicycle, and revels in the prospect, desires to have or ride the bicycle. And her revelling is not a symptom or concomitant—it *is* the desire. If we take away the revelling, both the real revelling and the disposition to revel, we take away the desire. Various complications would have to be addressed for this idea to turn into a viable explication of desire, but we will skip the niceties and focus on the core idea: positive or negative feelings about things are desires for or against them.

But then so are the moral sentiments, since they are one species of positive or negative feelings about things. This opens up a fourth way in which the sentiments can support the good. They can be desires and can as such give the person who harbours them reasons for doing the right thing.

From a utilitarian point of view, they can do so either more directly or less directly, depending on how they relate to our considerations from sections 3 and 5. If the sentiments themselves are of the type discussed in the previous section—in other words, if they themselves are utilitarian—they can amount to desires that happiness be maximized. Such desires support utilitarianism directly. How they do so is one of John Stuart Mill's topics. Mill spends one chapter of his treatise on utilitarianism inquiring into the "motives to obey" the precepts of that doctrine, developing a theory that hinges on the "social feelings of mankind".[23]

The indirect route involves states of affairs that have little or no logical connection with maximizing utility. Think, for instance, of the state of affairs that certain criminals are punished. Suppose that an agent's moral sentiment in favour of such a state of affairs qualifies, along the lines explained in this section, as a desire that the state obtain. In a decision-situation in which the outcomes with the most utility are the outcomes in which the state obtains, and the ones with less utility are not, this non-utilitarian desire will give the agent a sterling indirect reason to do what utilitarianism requires to be done.

We speak of a *sterling* indirect reason here in order to emphasize that such a reason involves a full-blown sentiment and thus a full-blown desire (for instance, that criminals be punished). This time, we are not dealing with a quick and dirty non-utilitarian sentiment that would evaporate under full representation, and not with a disposition that is cultivated merely as a

decision-making device by a sophisticated utilitarian who is trying to proceed indirectly. Impure states like these exist and have their point, and we have first encountered them in section 3.

But in that section we also considered the possibility that for some or even many people the non-utilitarian sentiments are and remain the axiological bottom line. We mentioned Adam Smith, who argues that, although we ultimately approve or disapprove of qualities in view of other features than their utility or hurtfulness, we thereby find ourselves approving of qualities that *are* useful and disapproving of qualities that *are* hurtful.[24] If Smith is correct, it should be possible to extend his observation from sentiments concerning qualities of mind to sentiments concerning actions or results. The sterling indirect reasons would be reasons involving the sentiments, and thus desires, that confirm to that pattern. Such an indirect reason for an action that meets the demands of utilitarianism holds for people not in as far as they are short of resources to turn utilitarian desires into rational decisions or actions, but in as far as they are short of utilitarian desires.

7. Changing the Payoffs: On the Difference between Altruists and Refined Egoists

Moral sentiments, we have argued, can serve or colour the utilitarian good in numerous ways: as tools, constituents, definers, evidence, and, directly or indirectly, as desires. They can also provide incentives for it. They can tie in with the agent's own happiness or unhappiness in ways that make the option that is better by utilitarian standards more attractive to her.

Consider the situation sketched in the grey area of Figure 1 (following page). You own a store, and the manager who runs it for you could cheat you. If the manager has no moral sentiments that tell against cheating, and does indeed cheat, she will gain $500 (by earning $1500 instead of $1000) and you will lose $1500 (by 'earning' –$500 instead of $1000). The payoffs are given in dollars but could just as well be given in 'units of happiness', and our utilitarian discussion of the financial example will indeed equate dollars with hedons. But now let us modify the manager. Let us suppose that the manager would feel bad about cheating after all, and that the bad feelings would mar the fun of the loot. Suppose that, if we convert her unhappiness into dollars, then all things considered (loot of $500 dollars, but with the hedonic loss) cheating would make her $9,500 worse off than not cheating.

This simple case illustrates how moral sentiments can change outcomes so that the outcome that is better by utilitarian lights becomes better for the

agent. The outcome that is better by utilitarian lights is in both cases the honest one, with a sum total of $2,000 (versus $1,000 and –$9,000, respectively). That better outcome is less attractive than the dishonest outcome for the manager without the qualms, but more attractive than the dishonest outcome for the manager with the qualms. The sentiments have pushed the utilitarian agenda yet again. And they did so without being directed at utility: they were sentiments against cheating and for honesty.

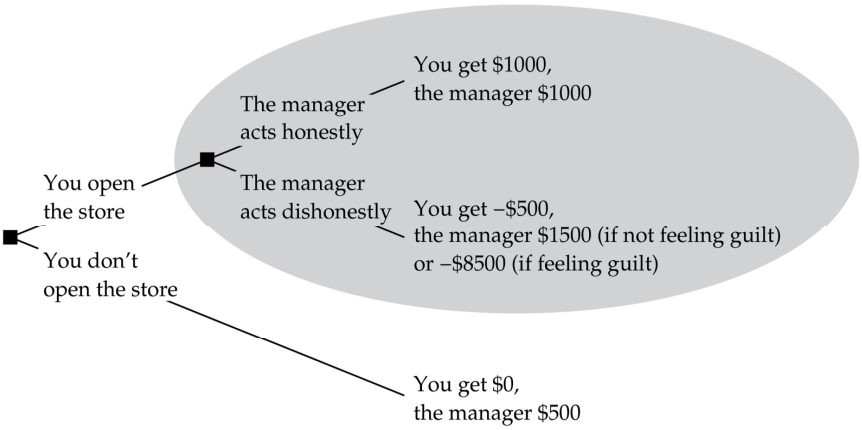

Figure 1: You and the manager

How does this pushing differ from that discussed in the previous section, where we pointed out that moral sentiments can be desires that give reasons to do the right thing? That is a legitimate question—not least because the drawing from Figure 1 could be read as illustrating the point from either of the two sections. And a culture has spread in the social sciences of reading such drawings sometimes one way and sometimes the other, with little awareness of the difference.

One way to read drawings involving options and payoffs is from preferences to welfare. You start with preferences between, or desires for, outcomes and explain the payoffs as numbers that reflect the strength of these preferences or desires. The other way—the way that this section is about—is from welfare to preferences. You start asking how happy an agent would be in the various outcomes or how good they would be for her, and explain the payoffs as numbers that reflect those amounts of happiness or

welfare. You work your way to preferences and reasons for action from there, by assuming that the person will desire or prefer the outcomes in proportion to the amounts of happiness or welfare they have in store for her.

In the class of cases that we are looking at, the difference between these two interpretations amounts, roughly, to the difference between genuine morality or altruism on the one hand and refined egoism on the other. The point from the previous section was that a moral sentiment can function as an intrinsic desire, say, to be honest or not to harm others. Such a desire can pull its utilitarian weight in decisions without any belief on the desirer's part that she herself will be happier after having refrained from cheating or harming another—happier, say, because she will rejoice ex post in her virtue or would otherwise suffer pangs of conscience. Whereas this section is precisely about such beliefs—about *their* role in making it wise for a person to do the utilitarian thing.[25] One utilitarian author who keeps drawing attention to the rational potential of these self-regarding expectations, too, is Peter Singer. He points out how frequently the individual who puts hard work into a worthy cause is rewarded with the joy of experiencing her life as fulfilled and meaningful.[26] In their capacity as an agent's predictable pleasures and pains, moral sentiments can provide, over and above moral reasons, self-interested reasons to work for the general good.

8. Solving Commitment Problems: New Payoffs and How to Signal Them

Let us extend our attention from the manager's decisions to yours and to their mutual dependence—that is, from the grey area to the rest of Figure 1. We now look at you in the stage in which you have to decide whether to open the store in the first place. At this stage the manager is merely the possible manager, and she happens to be the only candidate available.

If you open the store, you will enter the grey area that we have already discussed. If you don't open the store, your own income will be zero, which is better for you than the store with the dishonest manager but worse than the store with the honest manager; and the manager's income will be less than if she ran the store, either dishonestly without scruples or honestly.

Suppose that the manager has no scruples. If you open the store, she finds herself on the top branch of the decision tree, where she must choose whether to cheat. Believing the payoffs to be what they are, you predict

that the manager would cheat, which means your payoff would be –$500. And since that is worse than the payoff of zero you would get if you didn't open the store, your best bet is not to open it. This means a loss to both you and the manager relative to what could have been achieved had you opened the store and had the manager run it honestly. You two as well as every utilitarian onlooker will bemoan this waste of happiness.

The situation changes if the manager has, and you can find out that she has, the relevant scruples. We have already seen, in the grey area, how the hedonic burden of cheating transforms her payoffs. She becomes trustworthy. You can now be confident that, if you open the store, she will choose to manage honestly so that both you and she will come out ahead. The manager's conscience has given you a reason to open the store.

Notice the new dimension here. The previous section revolved around the decision of only one person, and the person's sentiments changed the payoffs so that she found herself with reasons to do what maximizes the public good. This time the sentiments do all that and more: they first bring about the very opportunity for the person to make that choice. They do so by creating what Tom Schelling calls a commitment.[27] The sentiments commit the manager to—that is to say, they give her a reason for—responding in a certain way to a certain action of yours, and this commitment of hers bears on your decision. By making it visibly wise for the manager not to short-change you if you create an opportunity for a joint gain, her sentiments make it wise for you to create that opportunity.

The visibility of the commitment is essential. A candidate for the post of manager who has the relevant moral sentiments without convincing you that she has them would never get a chance. With the epistemic part of the story being indispensable, it is worth asking how this sophisticated two-part affair—the sentiments plus the signalling-and-detecting that is required for them to function as commitment devices—can ever get off the ground. Signalling and its role in cooperation have attracted a good deal of attention in evolutionary biology and the behavioural sciences,[28] but the connection to the sentiments does not always receive its fair share of that attention.

One part of our question is how a signal of trustworthiness can emerge. Even if the first trustworthy person had borne some observable marker (say, the letter 't' on her forehead), no one else would have had any idea what it meant. Nico Tinbergen argued that a signal of any trait must originate completely by happenstance.[29] That is, if a trait is accompanied by an ob-

servable marker, the link between the trait and the marker had to have originated by chance. For example, the dung beetle escapes predators by resembling the dung on which it feeds. How did it get to look like this? Unless it just happened to look enough like a fragment of dung to have fooled at least the most dim-sighted predator, the first step toward a more dunglike appearance couldn't have been favoured by natural selection.[30] That first step must be a purely accidental link between appearance and surroundings. But once such a link exists, then selection can begin to shape appearance systematically.

Similarly, we may ask how a moral sentiment could have emerged if no one initially knew the significance of its accompanying marker. One hypothesis is suggested by the logic of the iterated prisoner's dilemma. There is no difficulty explaining why a self-interested person would cooperate in a prisoner's dilemma that is iterated with no end in sight. If you are a tit-for-tat player, for example, and happen to pair with another such player on the first round, you and that other player will enjoy the fruits of a long string of mutual cooperation.[31] For this reason, even Attila the Hun, lacking any moral sentiments, would want to cooperate on the first move of a prisoner's dilemma that has a sufficiently high probability to go on and on. However, you know that, if you cooperate on the first move, you forgo some gain in the present moment, since defection on any iteration always yields a higher payoff than cooperation. It is well known that both humans and other animals tend to favour small immediate rewards over even much larger long-term rewards.[32] So, even though you expect to more than recoup the immediate sacrifice associated with cooperation, you may discount those future gains excessively. Successful cooperation, in short, requires self-control.

If you were endowed with a moral sentiment that made you feel bad when you cheated your partner, even if no one could see that you had that sentiment, this would make you better able to resist the temptation to cheat in the first round. And that, in turn, would enable you to generate a reputation for being a cooperative person, which would be clearly to your advantage.

Moral sentiments may thus have originated as impulse-control devices.[33] The activation of these sentiments, like other forms of brain activation, may be accompanied by involuntary external symptoms that are observable. If so, the observable symptoms over time could have become associated in others' minds with the presence of these moral sentiments. And once that association was recognized, the sentiments would be able to play a second role—namely, that of helping people solve one-shot prisoner's dilemmas and other com-

mitment problems. The symptoms themselves can then be further refined by natural selection because of their capacity to help identify reliable partners in one-shot dilemmas.

The kind of signalling the foregoing account relies on, signalling that involves involuntary external symptoms, is well known in nature. Suppose, for example, that a toad meets a rival and both want the same mate. Among animals generally, the smaller of two rivals defers to the larger, thereby avoiding a costly fight that he would have been likely to lose anyway. Rival toads, however, often encounter one another at night, making visual assessment difficult. What they do is croak at one another, and the toad with the higher-pitched croak defers. The idea is that, on average, the lower your croak, the bigger you are. So it is prudent to defer to the lower croaker. This example illustrates the costly-to-fake principle: 'I'll believe you not because you *say* you are a big toad, but rather because you are using a signal that is difficult to present unless you really *are* a big toad.'[34]

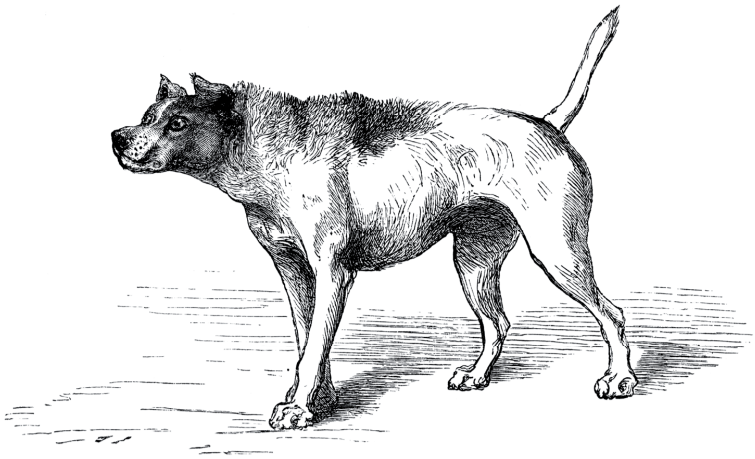

Figure 2: Dog approaching another dog with hostile intentions, drawn by Briton Riviere and included as Figure 5 in Darwin 1872.

It is the same when dogs face off: they follow an algorithm of deferring to the superior dog. Consider Figure 2, taken from Charles Darwin's book *The Expression of Emotion in Man and Animals*. We see a dog that is confronting a rival and is trying to establish its superiority. Darwin argued that we reliably infer what is going on emotionally in this dog by observing the numerous elements of its posture that are so serviceable in the combat

mode: the dog's hackles are raised, thus making "the animal appear larger and more frightful";[35] its fangs are bared, its ears pricked, its eyes wide open and alert, its body poised to spring forward. Any dog that had to go through a checklist to manifest these postural elements one by one would be too slow on the draw to compete effectively against a rival in whom the entire process was activated autonomously by the relevant emotional arousal. The autonomous link provides a window into the dog's brain.

Darwin mentions again and again that, for animals and humans alike, hereditary factors, instinct, reflex, and the force of habit will do more of the expressive work than conscious decision. Expressions tend to visit and leave our faces unbidden, and even in situations in which time is not the issue a deliberate attempt to cause or prevent certain expressions would fail.[36] This is part of what makes the expressions so telling.

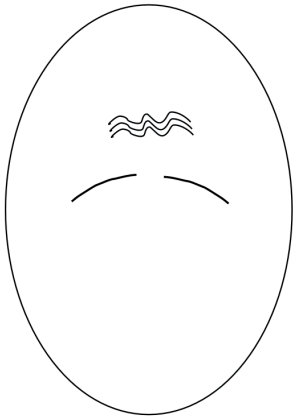

Figure 3: The characteristic expression of sadness or concern

An example is shown in Figure 3. People raised in different cultural traditions around the world can readily identify in this schematic portrait an expression of emotions like sadness or concern. As Paul Ekman and his colleagues have shown, most people are unable to reproduce this expression on command.[37] Various other emotions also have their characteristic signatures.

Signalling may be at work in more places than we think. Carl Bergstrom, Ben Kerr, and Michael Lachmann, for example, argue that a person's willingness to 'waste' time in social relationships may serve as a commitment device. According to their account, 'wasting' time would be a relatively costly step for defectors, who would be forced to seek other relationships anew

if discovered cheating.[38] The argument that lies behind us suggests a complementary interpretation: an inclination to spend seemingly unproductive time in social relationships may also be productive because it signals the presence of moral sentiments that also make cheating more costly.

As the reference to the role of cultivating social relationships suggests, the signalling and detecting in its entirety is a complex dance that plays out among people over time.[39] Dennis Regan, Tom Gilovich, and Robert Frank have done some experiments in this area.[40] The subjects in these experiments had conversations in groups of three for 30 minutes, at the end of which time they played prisoner's dilemma games with each of their conversation partners. Subjects were sent to separate rooms to fill out forms on which they indicated, for each partner, whether they were going to cooperate or defect. They also recorded their predictions of what each partner would do when playing with them. Each subject's payoff was then calculated as the sum of the payoffs from the relevant cells of the two games, plus a random term, so no one knew after the fact who had done what.

Almost 74 percent of the people cooperated in these pure one-shot prisoner's dilemmas. This finding, although it challenges economists' standard take on rational choice, is not unprecedented; other empirical studies have also found high cooperation rates in dilemmas when subjects were allowed to communicate.[41] The specific question of this study, however, was whether subjects could predict how each of their specific partners would play. When someone predicted that a partner would cooperate there was an 81 percent likelihood of cooperation (as opposed to the 74 percent base rate). On the defection side, the base rate was just over 26 percent, but partners who were predicted to defect had a defection rate of almost 57 percent. This seems an astonishingly good prediction on the basis of just 30 minutes of informal conversation.

The coevolution of sentiment-based trustworthiness and its recognition deserves to be explored in far more detail, but we are beginning to understand how this tandem can emerge, work, and persist. Most importantly for the purposes of this paper, we have evidence *that* it exists: trustworthiness has a considerable chance of being recognized. It follows that the moral sentiments that make people trustworthy have a considerable chance of solving commitment problems and of promoting in this way, too, the common good.

9. Conclusion

The real or alleged discords between moral feelings and utilitarianism preoccupy more philosophers and psychologists than the concords. We set

out to counterbalance this lopsidedness. Drawing partly on previous observations by other writers, we have identified and distinguished a variety of ways in which moral feelings beckon us towards the greatest happiness of the greatest numbers.

Some of these ways involve moral sentiments that are non-utilitarian or even anti-utilitarian. For one thing, such sentiments have long been appreciated and enlisted by the utilitarian tradition as unwilling accomplices in the maximizing of happiness (section 3). More interestingly and more directly, they are also *parts* of the general happiness or unhappiness and as such will often assimilate—from within—the utilitarian assessment of a situation to their own (section 4).

But there is no reason to look at non-utilitarian sentiments only. In a sentimentalist metaethics, sentiments might well turn out to provide the utilitarian doctrine with its foundation, and us with knowledge of that fact (section 5). Furthermore, moral sentiments of either kind, non-utilitarian or utilitarian, will often give full-fledged reasons for actions that are right by utilitarian standards. They do so both as an agent's desires (section 6) and as sources or parts of an agent's pain and pleasure (section 7), giving reasons not revolving around the agent's self-interest in the one case and revolving around it in the other. In particular, the reasons they generate will often enable us to defuse commitment problems (section 8), which count among the nastiest and most pervasive impediments to human happiness.

Do moral feelings provide or support an adequate morality? Is utilitarianism an adequate morality? We have said very little about those two questions, but we have pointed out that the answers to them are likely to resemble each other more strongly than is usually thought.

Christoph Fehige

Universität des Saarlandes

Robert H. Frank

Cornell University

NOTES

1. Cf. Foot (1967/1978), p. 23, Thomson (1990), ch. 7. Petrinovich, O'Neill, Jorgensen (1993) and Greene et al. (2001) count among the psychological studies.

We thank Eva-Maria Engelen, Bernward Gesang, and Stephan Schlothfeldt for bibliographic advice concerning some of the issues we touch upon in this paper; Alexander

Görres for assistance with some of the sources; and Thomas Fehige-Lutz and Stephan Schweitzer for helping us with the drawings and scans.

2. Several cases are mentioned in Foot (1967/1978) and Thomson (1990). A utilitarian sheriff who is inclined to punish an innocent man (although without getting him killed) appears in McCloskey (1957), pp. 468f., and Jim the botanist, asked to shoot an innocent person, in Williams (1973), pp. 98f. For a utilitarian perspective on such cases, see Hare (1981), ch. 8 and secs. 3.2 and 9.7.

3. Sidgwick (1874/1907), §3.14.5, chs. 4.3–4.5, and §1 of the "Concluding Chapter"; see also Mill (1861/1969), pp. 224f., Spencer (1862/1904). Sidgwick's discussion of the relation between utilitarianism and its method on the one hand and the morality of common sense or moral intuitions or sentiments on the other extends over more than 80 pages and covers far more than the point that we are focusing on in this section. Remarkably, there are no trolleys, botanists, or potholers in Sidgwick. He does not bring up a single contrived case of the type that dominates later discussions and for which we gave several examples in section 1.

4. Simon (1955) is a classic; see also Simon (1957) and, for a later statement, Simon (1983), ch. 1; Gigerenzer (2006) is a helpful survey of Simon's and other approaches.

5. As does Charles Darwin—see sec. 8 of this paper. For later developments, see, e.g., Simon (1967), LeDoux (1996), Oatley, Jenkins (1996), ch. 9. Some thoughts on the practical value of emotion have found much resonance in the wider public. This holds true of Antonio Damasio's theory of emotions as mechanisms of decision-making (Damasio 1994, esp. chs. 7–9) and even more so of various people's work on "emotional intelligence", as summarized and popularized in Goleman (1995).

6. Hare (1981), esp. chs. 3 and 8. Other important treatments of levels and of similar ways out of self-defeat in practical reason include Railton (1984), Pettit, Brennan (1986), and Parfit (1984/1987), part 1.

7. The long quotation is from Sidgwick (1874/1907), beginning of §4.5.1; the expressions "unconsciously Utilitarian" and "unconscious Utilitarianism" come up many times, in §§4.3.1, 4.3.7, and elsewhere.

8. Smith (1759/1979), ch. 4.2 and the end of ch. 7.2.3.

9. Austin (1832/1861), pp. 101 and 38 (first and second quotation) and 45 (long quotation); for the importance of rules, see pp. 42–45 and 52.

10. Mill (1861/1969), ch. 1, our emphasis.

11. Mill (1861/1969), pp. 224f.

12. Birnbacher (1996), pp. 242–51, (1998), sec. 2, (1998/2006), pp. 299–301, 313f., (2003), sec. 5.3.3.2, Gesang (2003), ch. 2. A similar move is anticipated and criticized in Williams (1973), pp. 103f.

13. In sections 6 to 8 of this paper we take another look at one fraction of these moral pleasures and pains as well as desires, namely at the agent's own. That other look concentrates on the states in another respect: less in their capacity of making an action morally right, and more in their capacity of making a morally right action rational.

14. Cf. Williams (1973), pp. 103f.

15. Hare (1981), esp. pt. 1 and ch. 8.

16. Theories of ethics that, one way or the other, put the moral sentiments centre-stage have been developed by Francis Hutcheson, David Hume, and Adam Smith. For modern versions, see Gibbard (1990) or Blackburn (1998).

17. Hume (1751/1998), end of sec. 3. Adam Smith's utilitarian streak is of a different kind—see secs. 3 and 6 of this paper.

18. All these claims can be found in Hutcheson's moral *Inquiry*, from 1725. The three quotations are from art. 5.2, the end of art. 3.8, and the introduction; the claim about the general consensus is made and argued for in secs. 3 and 4.

19. This line of thought is developed more fully in Fehige (2004a). The "nothing but" clause is worth adding, because representing too much impugns the aboutness of a response no less than representing too little. Any response coloured, say, by the belief that pushing people from bridges is *normally* no good is ipso facto no candidate for a response to the abnormal situation at hand.

20. Our proposal resembles R.M. Hare's theory of two levels of moral thinking, with one difference: Hare envisages an unsentimental and a sentimental level (1981), pt. 1 and ch. 8; we suggest that both levels could be sentimental.

21. LeDoux (1996), esp. chs. 6 (e.g., p. 166) and 9.

22. This tradition is partly documented in Fehige 2001 and 2004b.

23. Mill (1861/1969), ch. 3, quotations from pp. 227 and 231.

24. Smith (1759/1979), ch. 4.2 and the end of ch. 7.2.3.

25. Brilliant early campaigns for minding this difference can be found in Hutcheson (1728), articles 1.3f., and Butler (1726/1749), pp. xxiv–xxxiii and sermon 11.

26. E.g., Singer (1994), chs. 10f.

27. Schelling (1960), *passim*. Frank (1988) treats the role of the moral sentiments as commitment devices in more detail.

28. See, for example, Brown, Palameta, Moore (2003), Smith, Bliege Bird (2005), and the numerous references given in both.

29. Tinbergen (1952).

30. Cf. Gould (1977), p. 104.

31. Axelrod (1984), esp. pt. 2, Frank (1988), chs. 2 and 4.

32. Ainslie (1992), ch. 3.

33. See also Smith (1759/1979), ch. 4.2.

34. More on toads and similar cases in Searcy and Nowicki (2005); pp. 169–78 and 215f.; see also Frank (1988), ch. 6.

35. Darwin (1872), p. 95; see also p. 61.

36. Darwin (1872), pp. 48f., 186f., and elsewhere.

37. Ekman (1985), ch. 5.

38. Bergstrom, Kerr, Lachman (2008).

39. For a rich description, see Sally (2000).

40. Frank, Gilovich, Regan (1993).

41. See Sally (1995).

REFERENCES

Ainslie, George 1992. *Picoeconomics*, Cambridge: Cambridge U. P.

Austin, John 1832/1861. *The Province of Jurisprudence Determined* (first ed. 1832), second ed., London: John Murray.

Axelrod, Robert 1984. *The Evolution of Cooperation*, New York: Basic Books.

Bergstrom, Carl, Ben Kerr, and Michael Lachmann 2008. "Building Trust by Wasting Time", in *Moral Markets*, ed. by Paul Zak, Princeton, NJ: Princeton U. P., 142–53.

Birnbacher, Dieter 1996."Ethische Probleme der Embryonenforschung", in *Fragen und Probleme einer medizinischen Ethik*, ed. by Jan P. Beckmann, Berlin: de Gruyter, 228–53.

——— 1998. "Praktische Ethik als ethische Pragmatik", in *The Role of Pragmatics in Contemporary Philosophy*, ed. by Paul Weingartner et al., Vienna: Hölder-Pichler-Tempsky, 336–51.
——— 1998/2006. "Aussichten eines Klons" (first publ. in 1998), in *id.*, *Bioethik zwischen Natur und Interesse*, Frankfurt: Suhrkamp, 294–314.
——— 2003. *Analytische Einführung in die Ethik*, Berlin: de Gruyter.
Blackburn, Simon 1998. *Ruling Passions*, Oxford: Oxford U. P.
Brown, William Michael, Boris Palameta, and Chris Moore 2003. "Are There Nonverbal Cues to Commitment? An Exploratory Study Using the Zero-Acquaintance Video Presentation Paradigm", *Evolutionary Psychology*, 1, 42–69.
Butler, Joseph 1726/1749. *Fifteen Sermons Preached at the Rolls Chapel* (first ed. 1726), fourth ed., London: printed for John and Paul Knapton.
Damasio, Antonio R. 1994. *Descartes' Error*, New York: G.P. Putnam's Sons.
Darwin, Charles 1872. *The Expression of the Emotions in Man and Animals*, London: Murray.
Ekman, Paul 1985. *Telling Lies*, New York: W.W. Norton.
Fehige, Christoph 2001. "Instrumentalism", in *Varieties of Practical Reasoning*, ed. by Elijah Millgram, Cambridge, MA: MIT Press, 49–76.
——— 2004a. *Soll ich?*, Stuttgart: Reclam.
——— 2004b. "Wunsch I", in *Historisches Wörterbuch der Philosophie*, vol. 12, ed. by J. Ritter et al., Basle: Schwabe, columns 1077–85.
Foot, Philippa 1967/1978. "The Problem of Abortion and the Doctrine of Double Effect" (first publ. in 1967), in *ead.*, *Virtues and Vices: And Other Essays in Moral Philosophy*, Oxford: Blackwell, 19–32.
Frank, Robert H. 1988. *Passions within Reason*, New York: W.W. Norton.
——— Thomas Gilovich, and Dennis Regan 1993. "The Evolution of One-Shot Cooperation", *Ethology and Sociobiology*, 14, 247–56.
Gesang, Bernward 2003. *Eine Verteidigung des Utilitarismus*, Stuttgart: Reclam.
Gibbard, Allan 1990. *Wise Choices, Apt Feelings*, Oxford: Oxford U.P.
Goleman, Daniel 1995. *Emotional Intelligence*, New York: Bantam Books.
Gould, Stephen Jay 1977. *Ever since Darwin*, New York: W.W. Norton.
Greene, Joshua D., R. Brian Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen 2001. "An fMRI Investigation of Emotional Engagement in Moral Judgment", *Science,* 293, 2105–108.
Hare, R.M. 1981. *Moral Thinking*, Oxford: Oxford U. P.
Hume, David 1751/1998. *An Enquiry Concerning the Principles of Morals,* (first ed. 1751), text based on the editions from 1772 and 1777, Oxford: Oxford U.P.
Hutcheson, Francis 1725. *An Inquiry concerning the Original of Our Ideas of Virtue or Moral Good*, the second treatise (pp. 99–276) of *id*., *An Inquiry into the Original of Our Ideas of Beauty and Virtue*, London: printed by J. Darby for W. and J. Smith et al.
——— 1728. *An Essay on the Nature and Conduct of the Passions and Affections*, the first treatise (pp. 1–203) of *id*., *An Essay on the Nature and Conduct of the Passions and Affections: With Illustrations on the Moral Sense*, London: printed by J. Darby and T. Browne for John Smith and William Bruce.
LeDoux, Joseph 1996. *The Emotional Brain*, New York: Simon and Schuster.
McCloskey, H.J. 1957. "An Examination of Restricted Utilitarianism", *Philosophical Review*, 66, 466–85.
Mill, John Stuart 1861/1969. *Utilitarianism* (first publ. in 1861), in *id.*, *Essays on Ethics, Religion and Society*, Toronto: University of Toronto Press.

Oatley, Keith, and Jennifer M. Jenkins 1996. *Understanding Emotions*, Oxford: Blackwell.
Parfit, Derek 1984/1987. *Reasons and Persons*, Oxford: Oxford U. P.; reprinted with minor corrections from the edition Oxford 1984.
Pettit, Philip, and Geoffrey Brennan 1986. "Restrictive Consequentialism", *Australasian Journal of Philosophy,* 64, 438–55.
Petrinovich, Lewis, Patricia O'Neill, and Matthew Jorgensen 1993. "An Empirical Study of Moral Intuitions: Toward an Evolutionary Ethics", *Journal of Personality and Social Psychology*, 64, 467–78.
Railton, Peter 1984. "Alienation, Consequentialism, and the Demands of Morality", *Philosophy and Public Affairs*, 13, 134–71.
Sally, David F. 1995. "Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958 to 1972", *Rationality and Society,* 7, 58–92.
——— 2000. "A General Theory of Sympathy, Mind-Reading, and Social Interaction: With an Application to the Prisoners' Dilemma", *Social Science Information*, 39, 567–634.
Schelling, Thomas C. 1960. *The Strategy of Conflict*, Cambridge, MA: Harvard U.P.
Searcy, William A., and Stephen Nowicki 2005. *The Evolution of Animal Communication*, Princeton, NJ: Princeton U.P.
Sidgwick, Henry 1874/1907. *The Methods of Ethics* (first ed. 1874), seventh ed., London: Macmillan.
Simon, Herbert A. 1957. Introduction to Part 4, "Rationality and Administrative Decision Making", of *id.*, *Models of Man*, New York: John Wiley and Sons.
——— 1955. "A Behavioral Model of Rational Choice", *Quarterly Journal of Economics*, 69, 99–118.
——— 1967. "Motivational and Emotional Controls of Cognition", *Psychological Review*, 74, 29–39.
——— 1983. *Reason in Human Affairs*, Oxford: Blackwell.
Singer, Peter 1994. *How Are We to Live?* London: Mandarin.
Smith, Adam 1759/1979. *The Theory of Moral Sentiments* (first ed. 1759), sixth ed. (from 1790), Oxford: Oxford U.P.; reprint of the Oxford 1976 edition, with minor corrections.
Smith, Eric A., and Rebecca Bliege Bird 2005. "Costly Signaling and Cooperative Behavior", in *Moral Sentiments and Material Interests*, ed. by Herbert Gintis et al., Cambridge, MA: MIT Press, 115–48.
Spencer, Herbert 1862/1904. Letter to John Stuart Mill clarifying Spencer's stance on utilitarianism, included in Spencer, *An Autobiography*, vol. 2, London: Williams and Norgate, pp. 88f.
Thomson, Judith Jarvis 1990. *The Realm of Rights*, Cambridge, MA: Harvard U.P.
Tinbergen, Niko 1952. "Derived Activities: Their Causation, Biological Significance, and Emancipation during Evolution", *Quarterly Review of Biology,* 27, 1–32.
Williams, Bernard 1973. "A Critique of Utilitarianism", in J.J.C. Smart and Bernard Williams, *Utilitarianism: For and Against*, Cambridge: Cambridge U. P., 77–150.