SELF-DECEPTION AND THE MORAL SELF

RICHARD HOLTON

Suppose that we are motivated by the moral judgments that others make about us: we want others to think well of us as moral beings. That may move us to act well. But equally, when we act badly, it may move us to deceive those who witness our transgression. We may deceive them in straightforward terms about what we did. Or, if that is not possible, we may deceive them about how to categorize our action, about our motivation in performing it, or about the knowledge that we possessed when we acted. We may say that this wasn't really a case of dishonesty but of tact; that we acted, not for any personal benefit that we happened to accrue, but for the benefit of others; or that we had no idea, when we acted, of any harm that our action might bring about.

Suppose, however, that another story is true: we are primarily motivated, not by how others judge us, but by how we morally judge ourselves. Suppose, that is, that we want to judge ourselves as morally good. Then again we may be motivated to act well; but again, when we do not, we may be motivated to deceive. Now, though, the deception will be self-deception. Deceiving oneself about what one has done may be hard, at least in the immediate aftermath when memories are clear. But deceiving oneself about such murky issues as how to classify one's actions, about one's motives, or about the prior evidence one had for certain outcomes, may be much easier.

This idea, that we are moved by wanting to see ourselves as good, and that we use self-deception to achieve it, is an old one; most pre-20th century discussions of self-deception were focussed on its moral importance.[1] More recent philosophical discussions of self-deception have tended to lose this moral focus, but it has remained at centre stage in much recent thought in a variety of disciplines. Some see the maintenance of our moral self-image as providing the essence of moral motivation (Bénabou and Tirole, 2011); others see self-deception as an essential but instrumental step in the deception of others (von Hippel and Trivers 2011). Whether or not we want to make such sweeping claims, we are certainly very accustomed to the idea that self-deception plays an important role in our moral lives: that even if we genuinely want to do the right thing, our parallel desire to maintain our moral self-image means that when we behave badly we will frequently fail to realize what we are doing.

[1] (Dyke 1614), often cited as the first work on self-deception, is actually more about self-ignorance. Something closer to the contemporary notion develops in the 17th century, and is refined through the 18th; highlights include Rochefoucauld, Nicole, Hobbes, Butler, Hume and Smith. For discussion see (Moriarty, 2011 ch8; Garrett, 2017). Note that for these thinkers the idea is not that we simply want to believe that we are doing the right thing; the stress is on our wanting to do the right thing, but being over-ready to believe that we are doing so when we are not.

The idea of the banality of evil embraces many themes, but a central one is that those who act badly convince themselves that what they do is not so bad. Roy Baumeister traces such an attitude on the part of many of those involved in the atrocities of the Twentieth century. For a compelling example—admittedly not at the most atrocious end—consider the results of Timothy Garton-Ash's quest, after the fall of the Berlin Wall, to interview those who had kept a Stasi file on him. Almost without fail he was met with a mixture of denial, minimization and self-justification. "What you find is less malice than human weakness, a vast anthology of human weakness. And when you talk to those involved, what you find is less deliberate dishonesty than our almost infinite capacity for self-deception." (Garton-Ash, 1997 p.223)

How well has this approach fared under psychological scrutiny? Our concerns here will be two-fold. The first is with the evidence that we do indeed go in for moral self-deception for some of the reasons sketched; or indeed, for other reasons. The second is with how this fits with the perennial issue of the nature of self-deception. A literal-minded approach models it on the deception of others: it holds that in central cases of self-deception we know the truth, but we succeed in hiding it from ourselves. On such an approach, self-deception would involve the simultaneous holding of contradictory beliefs, with a purposive manipulation of what is made available to consciousness. That immediately raises the problem of how it would be possible: how we can be at once clever enough to arrange the deception and then gullible enough to fall for it. An increasingly influential deflationary alternative holds that nothing like this is going on. There are two different ideas here. The first is that in self-deception the part that deceives doesn't have to be seen as a homunculus, a full-blown knowing agent with projects of its own (Johnston 1988). This is not so controversial now; in fact it is probable that even Freud, often seen as the originator of such an approach, didn't really see the unconscious self as anything like a separate second agent (Gardner 1993). More controversial is the idea that self-deceived agents need have no awareness, at any level, of the facts from which they are screening themselves. The alternative model here is the kind of self-serving bias that work in social psychology has shown enables us to persist in self-ignorance in many spheres (Mele 1997, Barnes 1997, Mele 2001). Such a process may be bad enough for our ordinary view of our moral selves (Doris 2015), but it certainly doesn't amount to anything like a purposive self-manipulation that is analogous to what is involved in the deceit of others.

I say that this second view is controversial, since the self-deception displayed in moral cases frequently looks to be *reactive:* it seem to involve a tactical tuning of response to any threat to the picture that we have of our moral self. Such a reactivity requires there to be some recognition of the threat. It can still be described as involving bias, but this is not a pre-existing bias. It is a bias that is created in response to the perception of the threat.

We start with the work of two economists, Bénabou and Tirole (2011). Their central idea is that moral behaviour is largely driven by self-signalling: we act morally to convince ourselves that we are moral, since our actions provide our primary source of information of what we are really like. Self-signalling—that is, behaviour that is motivated at least partly by a quest to form beliefs about oneself—is not particularly exotic, nor does it require self-deception: for instance, we routinely try things out to see if we like them (Bodner and Prelec 2002; Holton, 2016). And even in cases in which the behaviour is performed solely in order to show that one can do it, there may be nothing problematic. If I stand up straight in order to show to myself that I have good posture, that is one way of getting a good posture.

But in the case of moral behaviour things are less straightforward. For a start, unlike in the case of posture, motivation matters. We do not normally think of ourselves as acting morally primarily in order to form beliefs about our own moral rectitude; indeed it may be that such a motivation would be inconsistent with truly moral behaviour. Morality requires doing things for the right reasons, and trying to show to oneself that one is good is plausibly not amongst them.[2] So if this is my motivation in acting well, it had better not be clear to me that it is. I will need to be self-ignorant. More substantially, if I am acting well simply to convince myself that I am good, then any time that I can achieve that conviction without acting well—by avoiding challenging circumstances, or by telling a story that will put my actions in a better light—I am likely to take the easier course. Here it seems that I will need to move beyond self-ignorance to self-deception, for we might think that a more active policy would be needed to keep me ignorant throughout such manoeuvres. Quite what is involved for something to be self-deception is a question we shall return to in due course; for now, let us look at the alleged phenomena.

Bénabou and Tirole want to accommodate three kinds of finding; I treat them under their useful headings, fitting in other research along the way. In each case Bénabou and Tirole's argument is that the findings are best explained if we understand the agent as involved in self-signalling.

(i) *Unstable Altruism:* Rather than being robust across different circumstances, moral behaviour diverges radically in the face of apparently morally insignificant differences.

This is a fairly diverse class. Bénabou and Tirole cite findings that subjects are less likely to cheat if they are paid in cash rather than with tokens, or if they have read

[2] Of course, moral philosophers differ on quite how important this is, from Kant at one end who thinks that impure motivations destroy virtue, to Hume (*Treatise* I iv) at the other, who thinks that pride can buttress it.

the Ten Commandments or a university honour code before acting (Mazar, Amir, and Ariely 2008); they are more likely to steal a can of coke from a fridge than a dollar bill, and so on. Such behaviour might be explained as self-signalling: in these contexts the consequences for one's self-conception might be more salient, and less amenable to excuse. However, they might equally be explained by subjects wanting to *be* good, and not simply wanting to believe that they are: they may need reminding that this is what they want, or what it is that good behaviour requires. Bénabou and Tirole also cite the findings on moral credentialing, where earlier bad behaviour gives rise to a subsequent tendency to compensatory better later behaviour (e.g. Carlsmith and Gross 1969), and conversely earlier good behaviour licences worse behaviour later (Monin and Miller 2001; Mazar and Zhong 2010; Zhong, Ku, Lount, and Murnighan 2010). Again this is compatible with self-signalling, but it is also compatible with simply thinking that subjects want to be *good enough*. It is also complicated by converse findings from the cognitive dissonance literature that performing small good acts will subsequently make subjects more likely to perform larger good acts—the so-called 'foot in the door' effect (DeJong 1979; see Cooper 2007 for the materials to fit this into the current complexities of cognitive dissonance theory). Bénabou and Tirole aim to explain this discrepancy by saying that in these latter cases it is a weaker aspect of identity that is challenged, but they give no independent reason for thinking that, and the traditional cognitive dissonance explanation (once I have started to conceive of myself in a certain way I will tend to act in accordance with that conception) has strong support (though we have no account of quite how this is supposed to interact with moral credentialing).

Much more persuasive in support of the idea of self-signalling is the finding that subjects will seek to avoid information that could put them in a bad light, or will act in worse ways if they can seem to off-load some of the responsibility onto others. For instance, subjects in a 'dictator game' who can choose to allocate a sum of money equally between themselves and another ($5:$5), or to increase their share marginally at great cost to the other ($6:$1), will normally choose the more equal option. But now take a second game in which the share going to the subject is stated, but in which the share going to the second person is hidden, although it can be costlessly revealed by the subject. You would expect a subject who was genuinely concerned with behaving well to reveal that information before choosing how to act; but around half chose not to, opting for the greater benefit to themselves, while preserving their ignorance of the consequence for the second person. (Dana Weber and Kuang 2007; see also Lazear, Ulrike, and Weber 2009).

Such motivations can be overstated; in a further condition only around 25% of subjects showed what looked to be morally self-deceptive behaviour (Dana Weber and Kuang table 4, 'plausible deniability'); and in a different experiment, Dana, Cain and Dawes 2006 found that subjects were primarily concerned to deceive others, not themselves. So there are almost certainly varied motives here, and probably mixed motives within any one individual. Nevertheless, some people, in some circumstances, seem to be primarily motivated by self-signalling.

Other studies lend broad support to this picture. A relatively early US study (Gaertner 1973) looked for different levels of racial bias between Liberal and Conservative Party members in New York. Experimenters with identifiably white or black accents telephoned subjects, pretending to have broken down on a freeway, and to have dialled the wrong number while trying to contact a garage. Explaining that they had used up their last coin, they then asked for assistance in getting through to the garage. Gaertner found that liberals were more likely than conservatives to help black callers once the request had been made; but that they were more likely than conservatives to hang up on black callers before this point. Discussing the experiment, Miller and Monin (2016) suggest that liberal subjects were more likely to identify the situation that was evolving as a potential test of their moral self-image, and foreseeing the required behaviour as costly, they withdrew from it; conservative subjects, less concerned that maintaining their self-image would require them to help, were less likely to hang up. Here it seems that, in some subjects at least, moral self-signalling is playing an important role.[3]

Consider next studies on how much people are prepared to pay for things when they know that a proportion of what they pay goes to charity. In one study (Jung et al 2017) reusable bags were offered to shoppers outside a supermarket. Shoppers could choose how much they paid for the bag, but they could not choose what proportion of their payment went to charity—in different conditions this would be 0%, 1%, 50%, 99% or 100%. The move from 0% to 1% more than doubled the average amount paid, but further increases had very little effect. It seems that what matters most in determining what people are prepared to pay is whether there is something going to charity; how much matters far less. If they were primarily concerned with the benefit to the charity, that is odd. It makes more sense on a signalling model if the value of the signal is relatively coarse: that is, if the benefit to self-image is much the same however much the charity receives.

Suggestive findings also come from the much discussed phenomenon of 'crowding out', although here the issues are more complex. The central idea here is that adding a financial incentive for some behaviour can crowd out a prior moral motivation for it. The classic case for this was made by Titmuss (1970), who argued that paying for human blood, as in the US, would result in poorer quality blood than a purely voluntary system, as in the UK. Titmuss canvassed various arguments for this (for instance that payment would would encourage those with diseases to conceal

---

[3] Miller and Monin make a general distinction between situations that provide *opportunities* for self-signalling—which they gloss as those that could enhance the agent's self-image—and those that provide *tests*—those that could diminish it. Put like that the distinction surely doesn't partition: most cases will provide both possibilities of enhancing and of diminishing, depending on how the agent acts. Presumably the point is that the net effect on self-image can be compared to the cost of acting. Sometimes performing an expensive signalling act will bring only a small gain to self-image, whereas failing to perform it will bring a large reduction; situations involving such tests should be avoided by self-signallers. Conversely, sometimes a relatively cheap act will bring a large gain to self-image, and failing to perform it will bring a small reduction; situations involving such tests should be sought out. Others fall somewhere in between. The distinction is a nice one, but it is not clear that many of the cases discussed by Miller and Monin, with the plausible exception of the Gaertner's, really address it.

them), but central was the idea that a moral motivation to donate would be crowded out once payment was provided. This might seem surprising: you might think that if it is a good thing to give blood when you are not paid, it is still good to give it when you are. Here self-signalling might provide the explanation: if the aim is to show that you are morally motivated, then payment does greatly obscure this.

Titmuss's claims about blood provision in response to payment have been contested (his evidence was very thin), and they are still not fully clear, but a recent meta-analysis suggests that, at the very least, adding a financial incentive does not increase provision, which is itself contrary to standard economic models (Niza et al. 2013). Still, other explanations need to be excluded before we conclude that it provides evidence for self-signalling. One is that blood donation might provide signalling to others. Another, more radical, is that offering a financial inducement does not just change the information about motivation, but changes the agent's perception of the nature of the act itself: giving blood moves it from an act that is seen to fall within the moral sphere to one that is not. If that were the case, then there need not be any self-signalling involved: agents could be simply motivated to do the right thing, independently of any signal given. Various other findings do point in this direction. For instance, a famous Israeli childcare study found that adding a fine to discourage the late collection of children actually had the reverse effect: the explanation given was that parents came to see the fine as a fee that could be blamelessly paid, rather than understanding lateness as a moral issue. (Gneezy and Rustichini (2000); the framework comes from Fiske (1992); for experimental support see Heyman and Arieli (2004)). Strikingly, removal of the fines did not return the behaviour to the earlier level, a finding consistent with the 'intrinsic/extrinsic motivation' research, which finds that once someone moves to an extrinsic motivation (in this case, a non-moral one) it is hard to get back to an intrinsic one (Deci and Ryan 2000). This is hard to explain if there is *only* self-signalling going on: once the financial reward is removed, it should be clear that the motivation cannot be driven by it. But it is perfectly compatible with a combination of a self-signalling account with one that acknowledges this kind of moral categorization: it is only once an agent perceives an act as moral that performing it provides signalling information. More work is needed here to distinguish the various possibilities.

(ii) *Social and Antisocial Punishments*. Agents will punish others for not being moral enough, but equally they will punish them for being too moral.

A fairly extensive experimental literature indicates that subjects in trust games and the like are prepared to punish, even at cost to themselves, those who have behaved badly, even at cost to themselves (Fehr and). But this admiration of morality only goes so far. Take a familiar example: people who are vegetarian for moral reasons. Rather than holding them up as moral exemplars, non-vegetarians often treat them with a mixture of scorn and resentment. (Minson and Monin 2012) One could imagine various explanations for this. The non-vegetarians might genuinely disagree with the moral position; or if they have some secret sympathy with it, they

might be concerned that the vegetarians are raising the moral bar too high. But studies on this and similar cases suggest that a powerful factor here is again self-signalling. It is hard to maintain a view of oneself as morally good if it is clear that there are people who are morally better around; and an easier course than changing one's own behaviour is to derogate the standing of the would-be exemplars.

So, for instance, consider a case in which subjects are given a task to do that is itself morally worrisome: in the experiment they are asked to imagine that they were detectives faced with a burglary, whose job was to identify the most likely culprit from a set of suspects. The descriptions were designed so that far and away the most plausible culprit was the sole African American among the three suspects; almost all subjects dutifully followed the instructions and identified them as the culprit. They were then shown a response purportedly from another subject (a 'rebel') who, rather than identifying the African American, had written on the form "I refuse to make a choice here—this task is obviously biased. . . . Offensive to make black man the obvious suspect. I refuse to play this game." A second group did things the other way round: first they were given the rebel response to look at, and then they were asked to make the assessment themselves. Subjects in the second group, those who had not themselves acted before they saw the rebel response, were much more likely to judge the rebel to have acted in a more moral way than someone who played along with the task; in contrast, those who had already acted, and hence played along themselves, saw the rebel as no more moral. And when asked for comments, those in the second group tended to describe the rebel as 'strong-minded', 'independent' or suchlike; while those in the first group described them as 'self-righteous', 'defensive' and the like.

(iii) *Taboo Thoughts and Trade-offs*: There are certain thoughts that we judge it would be wrong even to entertain.

A final set of findings that Bénabou and Tirole invoke concern the unthinkable. A number of psychological studies have examined 'protected values', the violation of which people are reluctant even to contemplate: the price at which one would sell one's children, for instance (Tetlock et al 2000; Schoemaker and Tetlock 2012). There may be good reason to put limits on thinkability; it may well be that not being prepared to think about something is a good first defence against doing it (Williams 1973, 93–4; 1992). But reluctance here is certainly not understood in pragmatic terms. Rather, people who have transgressed against taboos on the thinkable tend to see themselves as having been corrupted, and to seek 'moral cleansing behaviour', such as performing other good tasks, in response.

It is possible still to see this as driven by a concern to be good: if the prohibition can be costlessly violated, it is not going to work very well. But There is also plausibility in seeing this as, at least partly, self-signalling behaviour: 'Good people would not normally have such thoughts; since I have had them, I had better do something to prove that they were anomalous'.

So taking these three sets of considerations together, there is good evidence that people are frequently in the business of moral self-signalling. There is plenty of evidence to think, contra Bénabou and Tirole, that that is not the whole story; but it part of it. And if it is to be effective, subjects had better not realize that this is what they are doing. If this is self-deception, our next task is to understand what it involves.

ACCOUNTS OF SELF-DECEPTION

A natural place to start on understanding self-deception is to model it on the deception of others. There is a predictably complex literature on the exact requirements for deception, but a reasonable starting point is that I deceive you if and only if I intentionally get you to believe something I know to be false. Such an account applied to self-deception brings us to the corresponding idea that people are self-deceived if and only if they intentionally get themselves believe things they know to be false. Yet that has been widely held as problematic both with respect to *process* and to *outcome* (see Mele 1997, where these are termed the *dynamic* and the *static* paradoxes respectively).

Taking outcome first, if I come to believe something I know to be false, then presumably I both believe it and disbelieve it, which, if not impossible, seems to involve a very radical failure indeed. That might be avoided by thinking that self-deception involves a *shift* in belief, so that what I once believed to be false I now, at my own hand, believe to be true. But that concentrates the problem at the level of process. For how can I at once be manipulative enough to engineer my own deception, and credulous enough to fall for it? It is not simply that I will need to change my mind on the subject matter of the deception itself; if the deception is to be successful I will have to arrive in a state of belief without realising how I put myself there.

In response, deflationary theorists want to understand self-deception along other lines, dropping the parallel with the deception of others. There are independent reasons for worrying about that parallel—deception of others is often achieved by speech, by straightforward lying, yet presumably no one achieves self-deception in that way. Self-deception is going to have to involve more subtle expedients involving the selection of evidence and the construction of rationalising hypotheses. Once we focus on them it becomes more plausible that self-deception can be achieved without believing contradictions, and without intentionally engineering one's own deception. A proponent of this approach is Al Mele, who argues that self-deception needs no more than the acquisition of a false belief as the result of the operation of a pre-existing motivated bias. More specifically, he wants to explain central cases of self-deception using what he calls the Frederich-Trope-Lieberman (FTL) model, according to which agents require greater evidence before they accept an aversive belief than they would to accept a sympathetic one (Mele 2001, 31ff).

To see how this might work, we'll consider two experiments that have received a fair bit of discussion, one by Quattrone and Tversky (1984), the other by Mijovic-Prelec and Prelec (2010). In the Quattrone and Tversky experiment, the subjects, who believed they were involved in a study on the effects of cold showers after exercise, were first asked to hold their forearms in a vat of iced water until they said that they would rather not tolerate it any longer. Then their pulse was taken and they were asked to exercise on a stationary bicycle, after which they were asked to repeat the iced-water test, again until they said that they would rather not tolerate it any longer. In each case subjects were made aware of how long they had kept their arms in the water. Crucially though, in the period between the two iced-water tests, subjects were given a mini-lecture on psychophysics, during which they were told (falsely!) that people fell into two broad groups, those with Type 1 hearts, with shorter life-expectancies, and those with Type 2 hearts, with longer. The distinction was allegedly revealed by the degree of tolerance shown for cold water after exercise. Some subjects were told that increased tolerance was a sign of a Type 1 heart, and hence of shorter life-expectancy, whereas others were told that it was a sign of a Type 2 heart, and hence of longer.

Quattrone and Tversky found that most subjects (around 70%) changed their tolerance in the second test relative to the first, in a way that gave them good news. That is, those who believed that increased tolerance was a sign of longer life-expectancy showed increased tolerance; whereas, conversely, those who thought increased tolerance was a sign of shorter life-expectancy showed decreased tolerance. When asked whether they had tried to shift their tolerance, the majority said that they had not. Those who denied that they had tried to shift were much more likely to conclude that they had the healthy Type 2 hearts than those who admitted that they had.

Does this show self-deception? Quattrone and Tversky were quite cautious in the conclusions they drew. They followed Gur and Sackeim (1979) in defining the self-deceived agent as someone with contradictory beliefs who engages in the motivated act of bringing the more favourable of these to their attention. They concluded that in this sense '[a] certain degree of self-deception was probably involved' (p. 247); though 'To be sure, self-deception and denial are not all-or-none matters. Even subjects who indicated no attempt to shift may have harbored a lingering doubt to the contrary' (p.243). We can summarize their conclusion as being that (i) most of their subjects were probably modifying their tolerance 'purposefully' to obtain a better diagnosis; that (ii) most to some degree both believed they were doing this and believed they were not; and that (iii) most were more aware of the second of these beliefs than of the first.

In some ways this experiment looks like a good parallel to the kinds of self-deceptive behaviour shown in the moral case: the subjects seem to be doing something to provide themselves with good news. Nevertheless, and even though it has been the focus of much discussion—Mele devotes several pages to explaining how the FTL model can explain it (Mele 2001, 85–91; Mele 2019)—there is something unsatisfactory about it. Most cases of self-deception involve a shift in

belief, or at least a shift from what the subject would have believed, without the deception, to what they believe with it. But in this case we have a shift in *desire*: in the second trial, the subjects want to remove their arms earlier or later than in the first trial, depending on the information they have. The only problematic belief in question is the belief about whether they have shifted their tolerance. Clearly here in many cases they are mistaken—they believe they have not shifted their tolerance when they have. But Quattrone and Tversky give no reason for thinking that they really believe they have shifted it. This looks less like self-deception and more like straightforward self-ignorance.

If we are to provide a proper test for the FTL model then, we need a case in which we really have good evidence to think that there is something more than self-ignorance going on. So let's move to the second experiment, by Mijovic-Prelec and Prelec, which does involve straightforward modification of beliefs to achieve self-signalling in what looks like a self-deceiving way. The experiment involved asking subjects, who knew no Korean, to classify 100 Korean characters as either 'male-like' or 'female-like'. More specifically, the subjects were asked to classify the characters on the basis of how they thought others would classify them given similar instructions. (The task is thus what Keynes called a 'beauty contest': the right answer is that which matches the majority opinion.[4])

In the first round, subjects were told that they would be rewarded with two cents for every classification that they got right. This was designed to give a baseline in which people were simply trying to do as well as they could. The second round was designed to provide a situation in which they might display self-deception. The central idea was to provide a more complex reward structure, but not to provide information about how well subjects were doing. Self-deception would be shown if subjects acted in ways that would in fact be irrational, but that could easily take to provide evidence that they were doing well.

The details of the second round were as follows: before characters were shown, subjects were asked to predict whether they would be male or female. Since no information was given, this would be a pure guess. Then the character was shown, and subjects were asked to determine the whether it was male or female, as in the first round. And as in the first round, subjects were rewarded with two cents every time they got it right, though in this round they got two cents for a correct guess, and two cents for a correct identification. In addition they were told that in this round there was a substantial bonus prize of $40 that would be awarded to subjects who did best. For this they were divided into two groups. One group was told that this would go to the three people who made the best guesses prior to the characters being displayed; call this the 'guess-bonus' group. In the other was told that it would go to the three who made the best assessments once they had seen them; call this the 'assessment-bonus' group.

---

[4] As with other such tasks that require apparently meaningless classifications (for instance Köhler's 'maluma'/'takete' task; for discussion see Styles and Gawne 2017) the authors found considerable convergence, between 60 and 65%.

Obviously the best strategy to maximize financial return would be to guess randomly (or to always predict one gender if one thought that was more highly represented), and then to make the most accurate assessment that one could when presented with the character. But recall that the subjects were getting no feedback on how well they were doing. They could however provide themselves with some apparent good news about the accuracy of their guesses if they skewed their assessments so that they tended to line up: if you have guessed that a character will be male, be more prepared to assess it as male when you get to see it. That of course will probably cost you money, since your assessments will be less accurate than they could have been; but it will provide you with some (short-term) good news. Moreover, the value of that good news would differ depending on which group you were in. If you were in the assessment-bonus group, where the bonus was offered for the greatest accuracy of assessment, then it would merely indicate that you would pick up more two cent rewards for lucky guesses, something that would not amount to very much—even if you got them all right, you would only win $2. But if you were in the guess-bonus group, where the bonus goes to the best guessers, the good news would be much more significant: it would show that you were more likely to win $40. So if the self-deception were motivated by the value of the good news, you would expect to see more of it in the first group than in the second.

That is exactly what Mijovic-Prelec and Prelec found. There are three ways in which a subject's assessments in the second round might diverge from their first round baseline assessments. They might diverge so that they systematically stand in line with the guesses; this would be providing good news about the guesses. They might diverge so that they systematically stand out of line with the guesses; this would be providing bad news about them. Or they might diverge equally in both directions; this would be providing no news either way. Mijovic-Prelec and Prelec found no subjects who gave themselves bad news; subjects were split between those who gave themselves good news, and those who gave themselves no news either way. Strikingly, the proportion giving themselves good news was much larger in the guess-bonus group—where the good news was more significant—than in the assessment-bonus group.[5]

How should we understand this case? We start with the self-deceived subjects' first-order judgments about the gender of the characters. Here the judgments clearly changed as a result of the changing reward structure, and presumably, the desire to get good news about the chance of winning the bonus. But there is no evidence that the judgments are reactive in the sense discussed above: no evidence that the subjects needed to make the unbiased judgment in order to be effective in making the biased one. Instead the first-order judgments look to be explicable very much along the lines of the FTL model. Once there was reason to want the character to be, say, female, then the evidence that it was female was given greater weight than the

[5] Measured at the 0.05 confidence level 73% of the guess-bonus group and 53% of the assessment-bonus group gave themselves good news; at the 0.001 level, this fell to 45% and 27% respectively. For a perspicuous representation, see the graph on p. 235.

evidence that it was male. Subjects didn't need to develop any new strategy for this case; a preexisting general purpose FTL strategy would do the job.

What of their beliefs at the second-order level? Presumably if they had known that they were skewing their assessments in this way, they would have failed to have delivered any good news. But there is no reason to think that they did; as with the Quattrone and Tversky experiment, this looks like simple ignorance. And the FTL approach looks to be able to explain other features of the case too. No subject altered all of their assessments to give themselves good news. Of the 80 subjects, only two showed a self-deceptive pattern in over 40% of trials; most of those who were self-deceiving kept it at between 20% and 40%, where the pattern would not have been so obvious. But even this looks to explicable: some subjects are simply more prone to the FTL effect than others; it reduces the tendency to believe what is unpalatable, but doesn't remove it altogether, so that even the strongly self-deceived are left with a broadly credible picture, especially where things are vague enough to allow for flexibility in interpretation (Sloman 2010).[6]

Our question is whether all the cases of moral self-deception can be explained in terms of the FTL approach or something like it. It certainly can seem as though there is some active manipulation going on *in response* to unwelcome knowledge, something that Gur and Sackeim tried to capture with the idea of an agent with simultaneous contradictory beliefs who engages in the motivated act of bringing the more favourable of these to attention. Perhaps there are features somewhat less stark than those, but which nonetheless amount to more than pre-existing bias, and which bring back the idea of a responsive process. There is, after all, a great deal of space between the idea of preexisting bias, and that of the intentional inducing of a contradictory belief; self-deception might sit somewhere in this space. Let's explore quite what such a space would look like.

To start we need to be clearer on what is really at issue between the deflationary account that Mele and others have championed and the account that sees self-deception as actively responsive. It will be helpful to step back from the details of the debate to see things a little more generally.

GOING BEYOND THE IDEA OF CONTRADICTION

Let's suppose that there is some property—the *bad property*—that I do not want to know is ever instantiated. It may be; it may not be; I simply do not want to know. Here are two naive strategies I might take:

---

[6] The account thus seems able to have something to say in response to the 'selectivity problem', which is concerned with the idea that an account needs to be able to explain how agents are selective in the self-deception that they exhibit (Bermudez 2003; Mele 2019). At least it can say something about differences in when the bias does and doesn't lead to belief. We will return to the issue of whether it can account for when they do and don't turn a blind eye at the end.

Blanket strategy I close my eyes to everything. I take in no new information whatsoever.

Reactive strategy I keep a careful watch on the world. Whenever I see that the bad property is instantiated, I turn my eyes away and vehemently deny that it is.

Clearly both of these strategies are problematic. The blanket strategy will do the job; since I take in no information whatsoever, *a fortiori* I take in no information that the bad property is instantiated. But for most people in most situations it is clearly far too strong: in keeping myself ignorant of the bad property, I keep myself ignorant of everything that I need to know. In particular, when the bad property is not instantiated, I won't have the good news that it isn't.

In contrast the reactive strategy is perfectly discriminating. I only close my eyes to cases where the bad property is instantiated, and maintain my knowledge of everything else. Its problem is the opposite. Knowledge cannot be so easily lost. Once I have seen the bad property is instantiated, no amount of avoidance and denial will undo my knowledge. If my denials are vigorous enough, I might come to believe them; but that will take me to contradiction rather than ignorance.

In response to these problems, either strategy might be refined. The blanket strategy might be made somewhat less blanket: I might refuse to look in certain preordained places, or give any credibility to certain sources of evidence. Or, when I do get evidence, I might weight it differently using certain preassigned criteria. The reactive strategy might involve less than full recognition of the bad property before I turn away: I might register it only unconsciously, or I might turn when my assessment of its likelihood is high enough. Nonetheless, the distinction between the two approaches is reasonably clear: in the first, I put in place a strategy that works without my needing to register the bad property in any way; in the second I register the bad property in some way, and then react on the basis of that.

I suggest that this distinction is what is centrally at stake in the debate between the deflationary approach and that which sees self-deception as involving reactive self-manipulation. Defenders of the deflationary approach see all self-deception as involving descendants of the blanket strategy. In contrast, those who think that the deflationary strategy cannot explain all cases of self-deception think that this is because some involve descendants of the reactive strategy. Their central idea is that it is the belief that something is the case, or at least the suspicion that it might be, that brings the self-deception on. It is exactly because I start to believe that things are bad—I form a certain *triggering* belief— that I come to self-deceptively believe that they are fine. I react to defend myself, but in order to do this I need to identify the threat, and identify the kind of response that would work. The simplest approach is to understand this in terms of contradictory beliefs—I continue to believe both the triggering belief, and a self-deceptive belief that contradicts it. But there are other, less extreme, ways of following that strategy.

A first complication is this: as Quattrone and Tversky point out, belief can be more or less certain. There is no contradiction in thinking that p is possible, and that not-p is also possible. But we do not escape something like contradiction just in virtue of having partial beliefs. If I think that p is very likely, and that not-p is also very likely, or that p is certain and that not-p is possible, then I may not be strictly contradicting myself, but I will be guilty of the probabilistic analogue: I will have violated the requirement of the probability calculus that the probability of p and the probability of not-p must sum to one. The self-deceived person will have something analogous to contradictory beliefs if they categorically maintain the belief that p, while thinking that not-p is a real possibility.[7]

A second complication: deception does not fundamentally concern individual propositions, but subject matters. This shows up in the grammar—we do not say that A was deceived *that p*, but rather that A was deceived *about some subject or topic* —but the issue goes deeper than that. If I tell you that my friend has gone overseas, when really he is hiding upstairs, I deceive you about where my friend is, but I also deceive you about a host of other things: about what I believe, about how many people there are in my house, about whether you will be able to vent your rage on my friend here and now, and so on.

Some of these further things will be strictly entailed by what I say, but they do not all need to be. If you ask whether a company is solvent and I, knowing that the receivers have just been called in, tell you quite truthfully that it has the highest possible credit rating, then I have certainly acted to deceive you. What I have said—that is has the highest possible credit rating—is consistent with the claim that it is not solvent; indeed, in this case both, for now, are true. But they are in tension, in the sense that believing the former would, in the absence of further information, naturally lead you to reject the latter. So deception can extend to other items in the relevant subject matter even when they are not entailed by what I say.

There is a parallel phenomenon in the case of self-deception. If I know that my son has been killed, but I convince myself that he is still alive, then I have contradictory beliefs. But if someone whom I would normally trust tells me that he has been killed, and I become all the more sure that he is still alive, my two beliefs—the triggering belief that A said he is dead; and my self-deceptive belief that he is alive —are not contradictory. Again though they are in tension, in the sense that, were it not for my self-deception, the triggering belief would have led me to the opposite conclusion.

Issues here are delicate though, for we need to to distinguish this from a deflationary approach. If I decide in advance not to believe in any talk about the health of my son, that is a blanket strategy, explained by the deflationary approach. If I hear talk that he is dead, and my self-deception is a response to my realisation

---

[7] I here skate over various issues about how we should understand partial belief, all out belief, and the like. My contention is simply that however we think of this, we have to find space for quasi-contradictory states along these lines.

that the talk is credible, then it is not. The crucial difference is whether I have a pre-existing blanket strategy for blocking certain types of inference or not. If I do, that is compatible with a blanket strategy; if I have to tune which inferences I make in the light of my evidence that the bad property maybe instantiated, that is not.

A third issue concerns the *timing* of the different beliefs that one might have. To believe a contradiction is to believe two contradictory things at once. Even in the second-person case, deception does not require the simultaneous holding of contradictory beliefs by the deceiver and the deceived: the deceiver might have forgotten what they once believed in the meanwhile; indeed, their deception may be all the more effective if they succeed in deceiving themselves alongside their victim (Trivers). All that matters in general is the causal influence of the deceiver's belief on that of the deceived; the contradiction may be temporally dissociated. But particular cases may require more. If I am planning an elaborate deception, one that requires constant manoeuvring in response to changing circumstances, I may well need to keep track of how things actually are as the deception unfolds. Suppose I decide, Iago-like, to convince you that your devoted lover is unfaithful. Getting your lover to protest their innocence, when I have contrived to stack the odds against them, is part of my plan, since it will make them appear all the more duplicitous; but my confidence that they will protest is grounded in my knowledge that they are innocent. If for some reason I come to believe that they will not protest, I will need to change my plan. Here then the successful execution of the deception requires an ongoing awareness of relevant facts about the subject matter about which I am deceiving you, ongoing in that they continue while the deception is operative.

The facts about timing are similar in the case of self-deception if we understand it as reactive in the way sketched above. Again there is no general need for the agent to simultaneously hold contradictory beliefs. What is needed, on the reactive model, is the casual influence of the triggering belief on the ensuing, conflicting, self-deceptive belief; we can think of that as giving rise to something approaching an extended contradiction, even if there is never a simultaneous one.

Is there an analogue, in the case of self-deception, to the need for an ongoing awareness of how things actually stand on the part of the deceiver? It is easy to sketch one (though recall that we are not at this point asking whether such a thing really happens). Suppose that I want to maintain a good impression of myself, and I suppose that I do this by filtering the information that comes to me. The flattering information I attend to; the derogatory I ignore. How do I know what is flattering and what is not? It could be that it is marked in some independent way that enables a prior filter: information that is flattering is likely to come just from these sources, so to them I attend. But I may be living in an environment with no such useful indicators. Then I will need to attend to each piece of information closely enough to see whether it is flattering or not. I will need, in an ongoing way, to know the truth in order to self-deceive.

To summarize then: while avoiding straight-out contradiction, agents engaged in reactive self-deception might have beliefs (or partial beliefs) that are in probabilistic

tension; beliefs that are in tension within a subject matter; and beliefs that are in tension over time, either in a one-off, or an ongoing, way. And these three of course can combine. Rather than spelling out all of the possibilities, I will speak broadly of a triggering state that is, by the agent's own lights, *in tension* with the self-deceptive belief, adding details as need be. Call this a *tension-trigger*. If Mele is right that states of self-deception result from bias, he will still think that they are triggered. But if his account is to avoid these weaker forms of contradiction, if he wants to keep them broadly in the blanket camp, he will not want to accept that they are tension-triggered, since he will not want the subject to recognize the triggers to be in tension.

We can now sketch three different possible types of mechanism. The first is the only sort of mechanism that a pure motivated bias account, following the blanket strategy and eschewing any kind of contradictory belief, could countenance:

> (i) *No tension-trigger*: the state of self-deception does not involve any triggering state that is in tension with it.

So, for instance I might be born with a tendency to discount the critical remarks of others. If this bias is to count as motivated, there will presumably be a beneficial defensive explanation for it, but the doesn't proceed by means of a defensive reaction to the realization that others are thinking badly of me.

In contrast, the self-deceived agent might need to keep track, in an on-going way, of the very facts that are in tension with those that they are deceiving themselves about—the first-person analogue of the Iago strategy described above:

> (ii) *Running tension-trigger*: maintaining the state of self-deception requires the constant monitoring of triggering states that are in tension with it.

In between these two we have a mixed strategy. Here the tension-trigger provides the cue to put a strategy in place, and influences the nature of that strategy; but the strategy itself is a local blanket strategy, not requiring ongoing monitoring of the trigger:

> (iii) *Up-front tension-trigger*: the state of self-deception does involve a triggering state that is in tension with it, but the triggering state need only be experienced before the self-deception takes place, and so does not need to be maintained through it.

So, for instance, a certain source of information might be identified as providing bad news, which results in a blanket decision not to monitor that source. This might result in first order self-deception: the state whose recognition prompts putting the policy in place might be in tension with the first-order beliefs that the self-deception engenders: it is because I hear you telling me bad news that I resolve

to avoid you in the future. But the clash is likely to be more salient at the second-order level. In many situations putting an effective policy in place will require some careful thought; but if it is to be effective, that thought, and the policy that results, had better not be transparent.

Corresponding to these mechanisms I'll speak of trigger-free self-deception (i.e. the kind of biased based self-deception of which Mele talks); running self-deception, where the subject keeps track of the trigger; and up-front self-deception.

what kinds of mechanism are involved in moral self-deception as we actually see it?

The last section was highly theoretical: the aim was to show the different sorts of self-deception that might be possible. The focus in this section is empirical: what grounds do we have for thinking that any of these are actual? I take it that we have plenty of evidence of pre-existing bias; trigger-free self-deception is not in question. What is contentious is whether there are cases where there is a tension-trigger: cases of running self-deception, or, if not that, of up-front self-deception

Given the multiple interpretations available of any real-world example, it is only in controlled studies that we can hope for an answer; and even in such studies, it is hard to be sure that alternative interpretations are not available. Let us start with the more extreme case.

Running self-deception

There is evidence for running self-deception, but from cases that are in some way abnormal. I start with a striking one, but with two caveats: the subject was suffering from hemispatial neglect, and there was only one of him. The lead author was again Mijovic-Prelec (1994).

Hemispatial visual neglect is a not uncommon effect of strokes and other brain injuries. Patients are apparently unable to see objects in one side (typically the left-side) of the visual field. But the visual processing areas of the brain remain undamaged; the problem lies somewhere else. Quite what the problem is remains contentious, and will not be addressed here (see Robertson 2009 for a general introduction). What is important here is that the neglect is often not complete. In a famous example, a patient shown two pictures of houses whose right sides were identical but left sides differed in that one was on fire and the other wasn't, judged them to be the same, but expressed a preference to live in the one that was not burning (Marshall and Halligan, 1988; see Dorricchi and Galati 2000 for replication and development; and compare the similar phenomenon in Volpe *et al.* 1980).

FC, the subject in the Mijovic-Prelec study, was showing left-side visual neglect as the result of a stroke a month before. He was told that a dot might, or might not, be displayed on a screen in front of him; he was asked to say whether or not it was. There were three conditions: a dot on the right hand side; a dot on the left; or no dot. Unsurprisingly given his neglect, FC was able to correctly identify the presence of the dot on the right-hand side; able to correctly identify its absence; but normally unable to identify its presence on the left (he said it was absent). What was surprising was the reaction times. When the dot was present on the right-hand side his response was twice as fast as when it was absent: seeing the spot enabled him to stop a more laborious search. But when the spot was present on the left-hand side, his response—that is, his denial that a spot was present—was as fast as his recognition when the spot was presented on the right. It seems that at some level he saw the spot on left, which was enough to tell him that it wouldn't be present in the right-hand field that he could consciously see.

Is this self-deception? It doesn't fit a certain paradigm, in that it isn't motivated. But structurally it looks to be: in some way FC registered that the dot was there, and then he sincerely denied that it was. If so, this is clearly a case of running self-deception. There is no systematic bias that would enable FC to do what he was doing. Instead, he had to register, each time, that the dot was there on the left hand-side, in order to conclude so quickly that it was not.

Clearly this is just one case, but it does fit with other results from hemispatial neglect as mentioned above (see also Bisiach 1985). Stroke or other brain injury can give rise to other conditions that are naturally seen as self-deceptive. Neil Levy makes a plausible case for it in anosognosia, the denial of illness by those suffering from it, although the phenomena he reports look more like upfront self-deception than running (Levy 2009).

Even if these cases are widespread, there is an obvious concern that the afflictions facing those with brain injuries are hardly indicative of the capacities of those without. Perhaps that is right; but it would be somewhat surprising if brain injury, which typically and understandably depletes capacity, in this case generates a new one. What looks more likely is that there are mechanisms that normally keep distinct systems in line that are damaged here. If so, then it raises the possibility that the separate operation of those distinct systems could give rise to self-deception in the normal case. Moreover, cases other than brain damage can give rise to similar features. Patients suffering from visual conversion disorder (what was once known as hysterical blindness, and is often now called functional blindness) sincerely claim not to be able to see anything, but their visual systems are undamaged, and their behaviour on visual discrimination tasks differs from that of organically blind subjects, sometimes worse than chance, sometimes better (Bryant and McConkey, 1989, 1999).

Nevertheless, when we look for clear documented examples of running self-deception in otherwise normal subjects, none are obvious. It would be good to have experiments designed expressly to test it. In particular, none of the cases of moral

self-deception documented here seem to need it. That is not to say though that they can all be explained by the FTL; for there is reason to think that they require up-front self-deception. To this we now turn.

Up-front self-deception

The FTL model was based on the idea that self-deceived subjects received evidence that could have given them knowledge about how things really stood, but did not because of their biased belief forming practices. (Whether that is the best way of characterizing what is happening—whether we can distinguish knowledge and evidence in this way—is controversial (see, for instance, Williamson 1997), but presumably some sense can be made of the distinction.) But what about cases in which the subject avoids gaining evidence, presumably because, were they to get it, they would not be able to avoid forming the unwelcome belief. We saw this in the experiment by Dana, Weber and Kuang. Recall that there, subjects in a game were presented with a choice between an option which gave them $5 and a co-player $5, or an option which gave them $6 and the co-player $1. Most took the former option: they sacrificed $1 to significantly improve the other's lot. Other subjects, also told that they could choose between $5 or $6 for themselves, were told that the other player's share, again either $5 or $1, had been attached to one or the other of these options by a toss of coin (so that, for instance, choosing the $6 option might bring the other player $1, or, with an equal chance, $5). The other player's share was hidden, but could be revealed by the press of a button, and yet around half chose not to reveal it, taking the $6 without knowing what the other got. Or recall the Gaertner experiment in which liberal subjects, who were generally more likely to help out a black caller, were more likely to hang-up before any request could be made.

In these cases there is an up-front policy. Here it is a simple one, an easy blocking of a certain piece of information. But presumably in many real-world cases the policy will have to be rather more adaptive. Can it be achieved by the FTL approach? It seems not, for the subject will have to recognize, at some level, that a certain source of information will bear on the point at issue. In the Dana study, they do not know that revealing what the other gets will show them that taking $6 for themselves is wrong, but they know that it might show them that. That, it seems very plausible, is why they choose not to reveal it.

Would a prior blanket strategy enable them to decide which sources to ignore? Mele, in discussing a similar worry, suggests that people might ignore certain sources of information 'because they found expose to it very unpleasant' (Mele 2002, p. 48). That may be so, but why do they find it unpleasant? Because, in this case, they judge that it may give them information that would preclude them from taking the $6 and maintaining their moral image. But that involves acting on information which is itself in tension with the belief that they are trying to maintain: they want to believe that they are moral agents, with the openness to

relevant information that that requires, but they are now acting to avoid information that they know such a moral agent should seek. The FTL approach made a distinction between evidence and belief; this is not available here, since the agent needs to process the relevant evidence to know where to look and where to avoid.

If there is a way of blocking the idea that there are tension triggers at work here, it is by denying that there is false belief at the second-order level. In the Dana case, they must realize that they are avoiding information. So perhaps they are holding that that is perfectly compatible with being a moral agent. Here we need more information. It might seem surprising that someone might think that a certain course of action is precluded once it is known what it is, but that there is no requirement to get the information about it, even if it costs nothing. Yet that approach is far from absurd. Subjects may draw a doing/allowing distinction here: it is one thing to be guided by information that one has, another to require that one gets it. Likewise in the Gaertner experiment: one thing to refuse a request, another to wilfully make it impossible for the request to be made. Such distinctions may look like sophistry when clearly spelled out, but even if they are ultimately indefensible, this may indicate moral ignorance on the part of the subjects and not up-front self-deception.

In conclusion then, there remains much to do. We have plenty of evidence that moral behaviour is pervaded with self-deception; but discovering its nature will need more work.

BIBLIOGRAPHY

Baumeister, Roy, 1999. *Evil* (New York: Henry Holt)

Bénabou, Roland and Jean Tirole 2011. 'Identity, Morals and Taboos: Beliefs as Assets', *The Quarterly Journal of Economics*, 126, 805–55

Bermudez, Jose Luis 2003. 'Self-deception, intentions, and contradictory beliefs' *Analysis* 60, 309–319.

Bisiach, E., A. Berti and G. Vallar. 1985. 'Analogical and logical disorders underlying unilateral neglect of space'. In M. Posner and O. Marin (eds.) *Attention and Performance*, Vol . 11, (Hillsdale, NJ: Lawrence Erlbaum) 239-249

Bodner, Ronit, and Drazen Prelec 2002. 'Self-signaling and diagnostic utility in everyday decision making.' In I. Brocas and J. Carillo (eds.) *Collected Essays in Psychology and Economics* (New York: Oxford University Press) 105–126

Bryant, Richard, and Kevin McConkey, 1989. 'Visual conversion disorder: A case analysis of the influence of visual information', *Journal of Abnormal Psychology*, 98, 326–9

Bryant Richard and Kevin McConkey, 1999. 'Functional Blindness: A Construction of Cognitive and Social Influences', *Cognitive Neuropsychiatry*, 4, 227–41

Cooper, Joel, 2017. *Cognitive Dissonance: Fifty Years of a Classic Theory*. (London: Sage)

Dana, Jason, Daylian Cain, and Robyn Dawes, 2006. 'What You Don't Know Won't Hurt Me: Costly (but Quiet) Exit in a Dictator Game', *Organizational Behavior and Human Decision Processes*, 100, 193–201

Dana, Jason, Jason Xi Kuang, and Roberto Weber, 2007 'Exploiting Moral Wriggle Room: Experiments Demonstrating an Illusory Preference for Fairness', *Economic Theory*, 33 67–80

Deci, Edward, 1971. Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology*, 18, 105–115.

Deci, Edward and Richard Ryan, 2000. 'Self-Determination Theory' *American Psychologist* Vol. 55, 68–78

Doricchi, F and G. Galati 2000. 'Implicit Semantic Evaluation of Object Symmetry and Contralesional Visual Denial in A Case of Left Unilateral Neglect with Damage of The Dorsal Paraventricular White Matter', *Cortex* 36, 337–50

Doris, John, 2015. *Talking to Our Selves* (Oxford: Oxford University Press)

Dyke, Daniel 1614. *The Mystery of Selfe-Deceiving* (London: Griffin)

Fehr, Ernst and Urs Fischerbacher, 2003. 'The nature of human altruism', *Nature* 425, 785 –91

Fiske, A.P. 1992. 'The four elementary forms of sociality: Framework for a unified theory of social relations', *Psychological Review*, 99, 689–723.

Gaertner, Samuel 1973. 'Helping Behavior and Racial Discrimination among Liberals and Conservatives', *Journal of Personality and Social Psychology*, 25, 335–41

Gardner, Sebastian 1993. *Irrationality and the Philosophy of Psychoanalysis*. (Cambridge: Cambridge University Press)

Garrett, Aaron 2017. 'Self-Knowledge and Self-Deception in Modern Moral Philosophy'. In Ursula Renz, *Self-Knowledge: A History* (New York, OUP)

Garton-Ash, Timothy *The File*, (New York: HarperCollins 1997)

Gneezy, Uri, and Aldo Rustichini, 2000. 'A Fine is a Price', *Journal of Legal Studies*, 29, 1–18.

Gur, Ruben and Harold Sackeim 1979. 'Self-Deception: A Concept in Search of a Phenomenon', *Journal of Personality and Social Psychology* 37, 147–69

Heyman, J. and Daniel Ariely, 2004. 'Effort for Payment: A Tale of Two Markets', *Psychological Review*, 15, 787-793

Von Hippel, William, and Robert Trivers, 2011. 'The Evolution and Psychology of Self-Deception' *Behavioural and Brain Sciences*, 34, 1–56

Holton, Richard, 2016. 'Addiction, Self-signalling, and the Deep Self', *Mind and Language* 31, 300–13

Johnson, Mark 1988. 'Self-Deception and the Nature of Mind' in B. McLaughlin and A. Rorty (eds.) *Perspectives on Self-Deception* (Berkeley: University of California Press) 63–91

Jung, Minah, Leif Nelson, Uri Gneezy and Ayelet Gneezy, 2017. 'Signaling Virtue: Charitable Behavior Under Consumer Elective Pricing', *Marketing Science* 36, 187–94

Levy, Neil.2009. 'Self-Deception Without Thought Experiments' in T. Bayne and J. Fernández (eds.) *Delusion and Self-Deception* (New York: Psychology Press) 227–42

Marshall, J.C. and Halligan, P.W., 1988. 'Blindsight and insight in visuo-spatial neglect', *Nature* 336, 766–7

Mazar, Nina, On Amir, and Dan Ariely, 2008. 'The Dishonesty of Honest People: A Theory of Self-Concept Maintenance'. *Journal of Marketing Research*, 45, 633–4.

Mazar, Nina, and Chen-Bo Zhong, 2010. 'Do Green Products Make us Better People?', *Psychological Science*, 21, 494–8

Mele, Alfred 1997. 'Real Self-Deception', *Behavioral and Brain Sciences* 20, 91-102

Mele, Alfred, 2001. *Self-Deception Unmasked* (Princeton: Princeton University Press)

Mele, Alfred, 2019. 'Self-Deception and Selectivity' *Philosophical Studies*

Mijovic-Prelec, Danica Mijovic-Prelec, D., Shin, L. M., Chabris, C. F. & Kosslyn, S. M. 1994. 'When does 'no' really mean 'yes'? A case study in unilateral visual neglect', *Neuropsychologia* 32, 151–158.

Mijovic-Prelec, Danica, and Drazen Prelec 2010 'Self-deception as self-signalling: a model and experimental evidence' *Philosophical Transactions of the Royal Society B* 365, 227–40

Miller, Dale and Benoît Monin, 2016. 'Moral Opportunities Versus Moral Tests'. In Joseph Forgas, Lee Jussim, and Paul van Lange (eds.) *The Social Psychology of Morality* (New York : Routledge) 40–55.

Minson, Julia and Benoît Monin, 2012. 'Do-Gooder Derogation: Disparaging Morally Motivated Minorities to Defuse Anticipated Reproach' *Social Psychological and Personality Science* 3(2) 200–7

Monin, Benoît and Dale Miller, 2001. 'Moral Credentials and the Expression of Prejudice', *Journal of Personality and Social Psychology*, 81, 33–43

Monin, Benoît, Pamela Sawyer and Matthew Marquez 2008. 'The Rejection of Moral Rebels: Resenting Those Who Do the Right Thing' *Journal of Personality and Social Psychology* Vol. 95, 76–93

Michael Moriarty, 2011. *Disguised Vices: Theories of Virtue in Early Modern French Thought* (Oxford: Oxford University Press)

O'Conner, Kieran, and Benoît Monin, 2016. 'When Principled Deviance Becomes Moral Threat: Testing Alternative Mechanisms for the Rejection of Moral Rebels' ,*Group Processes & Intergroup Relations*, 19, 676–93

Quattrone, George and Amos Tversky 1984. 'Causal Versus Diagnostic Contingencies', *Journal of Personality and Social Psychology*, 46, 237–48

Robertson, Lynn, 2009. 'Spatial Deficits and Selective Attention' in Michael Gazzaniga (ed.) *The Cognitive Neurosciences* 4th Edition, (Cambridge MA: MIT Press) 269–80

Sloman, Steven, Philip Fernbach and York Hagmayer, 2010. 'Self-deception requires vagueness' *Cognition* 115, 268–81

Syles, Suzy and Lauren Gawne, 2017. 'When Does Maluma/Takete Fail? Two Key Failures and a Meta-Analysis Suggest That Phonology and Phonotactics Matter', *I-Perception*, 8

Tetlock, Philip, *et al.*, 2000. 'The Psychology of the Unthinkable: Taboo Trade-Offs, Forbidden Base Rates, and Heretical Counterfactuals', *Journal of Personality and Social Psychology* Vol. 78, 853–70

Volpe, Bruce, Joseph Ledoux and Michael Gazzaniga, 1980.'Information processing of visual stimuli in an "extinguished" field', *Nature*, 282, 722–4

Williams, Bernard, 1973. *Utilitarianism: For and Against* (Cambridge: Cambridge University Press)

Williams, Bernard, 1992.'Moral Incapacity', *Proceedings of the Aristotelian Society* 92 59–70

Williamson, Timothy 1997.'Knowledge as Evidence', *Mind*, 106, 717–42.

Zhong, Chen-Bo, Gillian Ku, Robert Lount, and J. Keith Murnighan, 2010. 'Compensatory Ethics,' *Journal of Business Ethics*, 92, 323–39.