

**It's not just what you do, but what's on your mind:  
A review of Kwame Anthony Appiah's "Experiments in Ethics"**

Liane Young and Rebecca Saxe

Department of Brain and Cognitive Sciences, MIT

What could science have to offer philosophy, on the topics of morality and ethics? Some philosophers fear, and some scientists seem to hope, that morality and ethics will turn out to be topics of "natural philosophy", like space and time. That is, apparently normative problems will turn out to be best addressed by empirical investigation of the natural world, and moral judgments will turn out to be true or false just like claims about the indivisibility of ether or the speed of light. To his credit, Kwame Anthony Appiah begins "Experiments in Ethics" by distancing himself from these hopes and fears. Without confusing an "is" for an "ought", Appiah suggests that science has much to offer and much to take away from theories of morality and ethics. With this we agree.

In particular, Appiah suspects that scientific results may undermine moral intuitions either indirectly, by invalidating our factual assumptions about the causes of human behavior, or directly, by undermining our confidence in the actual sources of our intuitions. One specific moral theory that Appiah defends is "virtue ethics", and, more broadly, the intuition that what matters, for morality and ethics, is who you are on the *inside*, and not just what - in the sense of your effects on the external world - you do. Do scientific results compel philosophers (and the rest of us) to abandon this intuition? "Experiments in Ethics" reflects an uneasy sense that they both do, and do not.

Again, we agree with Appiah here. Scientific results both do, and do not, force us to abandon the intuition that internal causes of human actions ought to matter to our moral judgments. However, the resolution we offer is quite different from Appiah's. We propose that Appiah's dilemma arises partly from an over-simplified conception of "internal causes" that is shared widely among both scientists and philosophers. By re-introducing the true richness of internal causes invoked in moral judgments, we hope to relax the tension between scientific and intuitive theories of human behavior, and thus to comfortably accommodate "internalist" moral judgments. Nevertheless, we do not propose to remove this tension altogether. In critical cases, scientific discoveries about human behavior *should* undermine our moral intuitions, by changing the way we evaluate both others and ourselves.

In closing, we offer a general answer to the question with which we began, about the relationship between science and ethics: science can undermine our commitment to specific intuitive moral judgments but cannot strengthen it. This resolution should be appealing to both scientists and philosophers. The science of human nature can constrain morality and ethics, but philosophy remains the keeper of positive normative knowledge.

### **How we are, how we ought to be**

One way scientific facts may undermine moral beliefs is by invalidating the factual assumptions that support those beliefs. Moral intuitions depend on, among other things, our intuitive theory of human behavior, why people do the things they do. "After all," Appiah notes, "'ought' implies 'can'"; we should not hold people to impossible standards

of behavior. As a consequence, Appiah turns first not to the scientists who conduct 'experiments in ethics', the moral psychologists, but to social psychologists who study, in some sense, the gap between scientific and intuitive accounts of the causes of human behavior.

At the core of our intuitive understanding of human behavior is an assumption that people are generally free to act however they wish, that people's observable behavior reflects often their unobservable dispositions. There are obvious exceptions: for example, if a person is in chains or held at gunpoint, literally or metaphorically. But in the vast majority of cases, observers expect that it is in the person's power to "just say no". If such freedom exists, then actions on the external world should reliably reflect internal inclinations. There is evidence that observers proceed on this very assumption: observers are willing and ready to infer the 'insides' of person in direct correspondence to his or her observable 'outsides' (Gilbert & Malone, 1995).

In contrast to this intuition, studies in social psychology over the past five decades have revealed the tremendous power of forces in the environment (the "situation") to shape people's behavior, while flying under the radar of observers. One important example is the power of authority and/or consensus. Without holding either a literal or metaphorical gun, leaders and groups elicit extreme compliance from people (e.g., Asch, 1951; Milgram, 1963; Ross, 1988; Zimbardo, 1973). Our intuition predicts that people may at any moment just refuse to follow their leader or group, if the costs are high or the consequences unpalatable. But our intuition is wrong. Empirical observation shows that they almost never do.

Milgram's famous experiments represent the most dramatic demonstration of the power of authority. (Oddly, Appiah does not discuss the Milgram experiments directly, though he does discuss related research). Participants were told to act as the 'teacher' in a memory experiment, and deliver electric shocks to the other participant (the 'learner') as punishment for his or her incorrect answers or failure to respond. On successive trials the shock level was increased on a dial. The highest levels of shock were marked 'extreme danger', and 'XXX'; the 'learner' responded to successively higher levels of shock by objecting, complaining about a bad heart, screaming in pain, and eventually slumping silently in the chair. Two findings of the Milgram experiment are important: (1) the majority of participants continued to deliver the shock, when instructed to do so, up to the highest levels; but (2) when asked beforehand, observers agreed that only the tiny minority of sadist participants would go so high.

How do these scientific results threaten to undermine moral philosophy? Appiah considers both a narrow and a broad impact. Narrowly, these results reveal a specific mistake in some intuitive moral judgments. The mistake is in neglecting the contribution of situational force to people's evil actions. Watching one person incomprehensibly harm and degrade another person, with his eyes wide open and no gun to his head, we automatically infer evil insides that correspond to the evil actions. We falsely believe that most people (including ourselves) would have acted differently in the same situation, that 'just saying no' was the easy and obvious response, and that only the tiny minority of sadists in the population would behave in such a horrific way (e.g., consider the Rwandan genocide). Our mistaken assumptions about human nature result in our moral condemnation of the whole person, the bad apple, without regard to the

surrounding situation. Recognizing that our assumptions are wrong must impact our moral judgments (and our ethical choices). In this, we agree completely with Appiah.

However Appiah fears that these results pose a threat to the moral theory he favors, virtue ethics, and to the intuition that what matters, for morality and ethics, is who you are and not just what you do. Here, we disagree with Appiah's assessment of both the problem and the solution.

### **A false dichotomy**

Appiah's problem, as we see it, arises from a widely accepted but over-simplified dichotomous model of behavior attribution. Historically, scientific and philosophical theories have assumed that observers attribute behavior to either "something about the person" or "something about the situation" (e.g., Heider, 1958). In social psychology, "something about the person" traditionally means stable, distinctive traits of the individual that are consistent across situations and predict behavior across situations; this is related to the philosophical idea of "character". So, for example, an observer could decide that Judy ate the dog food because (1) there was no other food around (situation) or (2) she doesn't have very discriminating taste (stable trait).

As described above, social psychologists have argued that whereas observers typically attribute behavior to stable traits, in reality, behavior is driven powerfully by situations. Some (though by no means all) social psychologists even claim that there is *nothing* that corresponds to the intuitive idea of internal 'traits' or, in philosophical terms, 'character'. Appiah provides a clear review of this literature on "the situationist threat".

Given the dichotomous model of behavior attribution, the discovery that behavior is not determined by stable traits implies that behavior cannot be attributed to “something about a person” at all. Appiah recognizes this, and a further implication. Philosophical accounts depend on a distinct but related dichotomy: moral judgments are supposed to depend on either character (internal traits) or consequences (external outcomes). Appiah himself appears to favor some form of “virtue ethics” according to which (among other things) moral judgments depend on character; what matters is not just what we do, but the kind of people we are.

Appiah therefore feels himself backed into a corner. We cannot require ourselves or others to be compassionate and courageous people if such traits don’t exist; that is, if the only thing that does exist is a kind of situation that brings out compassionate or courageous behavior. So social psychology appears to force the moral philosopher to abandon character as a standard for moral judgments, and therefore (given the dichotomy) to embrace consequentialism.

As a solution, Appiah resists the dichotomy: “Only a misguided theoretical parsimony would make us choose between considerations of character and considerations of consequence” (see also Casebeer, 2005). We offer an alternative response: only a misguided theoretical parsimony would *limit* us to considerations of character and considerations of consequence.

### **It’s the thought that counts**

The standard dichotomy between stable internal traits and ever-changing external situations neglects a major factor in intuitive moral judgments: mental states, like beliefs

and desires. Perhaps Judy ate the dog food because she *thought* it was canned tuna or because she *wanted* to taste it. Beliefs and desires straddle the traditional division, because they are internal to the person, but depend on the situation, change with the situation, and often make no sense outside of the situation (Malle, 2004). Mental state explanations appear in everyday attributions, i.e. Why was she late? Because she thought class started at noon. Note that such a mental state explanation arises independent of any explicit appeal to character or situation causes or their relation. Recognizing the central role of mental state explanations relieves some of the tension between social psychology, virtue ethics, and ordinary intuition. An especially entrenched intuition of ours is that our moral judgments of an action and an agent should reflect not just the outward effects of the action but what was going on inside the agent's mind at the time of the action. Beliefs and desires are 'insides' in this sense. And social psychological evidence concerning the power of the situation does nothing to undermine the existence (or importance) of transient mental states.

Our own research has directly investigated the relative importance of consequences versus mental states in intuitive moral judgment (Young, Cushman, Hauser, & Saxe, 2007). Participants read stories in which agents produced either a negative outcome (harm to another person) or a neutral outcome (no harm), based on the belief that by acting they would cause the negative outcome ("negative" belief) or the neutral outcome ("neutral" belief). Participants then judge whether the action was permissible (or in other cases, how much blame the actor deserves).

Here is an example story in the "neutral belief, negative outcome (accidental harm)" condition: "*Grace and her friend are taking a tour of a chemical plant. When*

*Grace goes over to the coffee machine to pour some coffee, Grace's friend asks for some sugar in hers. There is white powder in a container by the coffee. The container is labeled "sugar", so Grace believes that the white powder by the coffee is sugar left out by the kitchen staff. The white powder is actually a very toxic substance left behind by a scientist, and therefore deadly when ingested in any form. Grace puts the substance in her friend's coffee. Her friend drinks the coffee and dies."*

Across all of our studies using these materials, adult participants weighed the agent's belief more heavily than the action's consequences in their moral judgments. A simple metric of this effect is that our participants almost universally judge an attempted harm (negative belief, neutral outcome) as more blameworthy and more forbidden than an accidental harm (neutral belief, negative outcome). Fiery Cushman (2008) has pushed this line of work even further, directly comparing the roles of consequences, causation, beliefs and desires for different kinds of moral judgments (e.g., person, permissibility, and deserved blame and punishment). The agent's belief about whether his action would cause harm was the most important factor across the board, followed by the agent's desire to cause harm.

Forgiving accidental harms is non-trivial. Among adults, we have found evidence of substantial individual variability in blame assigned to protagonists in our accidental harm scenarios (Young & Saxe, submitted). In development, this pattern of moral judgments does not emerge until approximately age seven, surprisingly late in childhood. Five-year-old children are capable of reasoning about false beliefs; in the paradigmatic "false belief task", children predict that observers will look for a hidden object where they last saw the object, not in its true current location (e.g., Flavell, 1999;

Wellman, Cross, & Watson, 2001). However, these same children say that if a false belief led an observer to unknowingly and accidentally hurt another person (e.g., mistake poison for sugar), she is just as bad as if she had caused the harm intentionally (e.g., Piaget, 1965/1932). The ability to integrate beliefs into moral judgments then appears to be a distinct developmental achievement. Consistent with this idea, high functioning adults diagnosed with Asperger's Syndrome, who pass traditional false belief tasks, also fail to withhold blame in our accidental harm scenarios (Moran, Young, Lee, Gabrieli, & Saxe, in preparation).

For most healthy adults, though, beliefs and desires carry more moral weight than external consequences. In some cases, beliefs and desires even overwhelm other morally relevant external factors, like whether the person could have acted otherwise. Woolfolk, Doris, and Darley (2006) presented subjects with variations of one basic story: *"Bill discovers that his wife Susan and his best friend Frank have been involved in a love affair. All three are flying home from a group vacation on the same airplane. In one variation of the story, their plane is hijacked by a gang of ruthless kidnappers who surround the passengers with machine guns, and order Bill to shoot Frank in the head; otherwise, they will shoot Bill, Frank, and the other passengers. Bill recognizes the opportunity to kill his wife's lover and get away with it. He wants to kill Frank and does so. In another variation: Bill forgives Frank and Susan and is horrified when the situation arises but complies with the kidnappers' demand to kill Frank."* On average, observers rate Bill as more responsible for Frank's death, and the killing as more wrong, when Bill wanted to kill Frank, even though his desire played no role in causing the death.

Most of our own research has focused on the neural mechanisms that support belief-based moral judgements. Our results suggest that specific brain regions support a number of aspects of mental state reasoning for moral judgment, for example, the initial encoding of the agent's belief, the use and integration of the belief (with outcome information) for moral judgment, the spontaneous inference of the belief for moral judgment in the case that this information isn't explicitly provided, and even post-hoc reasoning about the belief to support a moral judgment (Young et al., 2007; Young & Saxe, 2008; Young & Saxe, in press; Kliemann, Young, Scholz, & Saxe, 2008). The most selective brain region appears to be a patch of cortex just above and behind the right ear: the right temporo-parietal junction, or RTPJ. Recruitment of this region during healthy adults' moral judgments is correlated with individual differences in the extent to which "neutral" beliefs are used to forgive accidental harms.

Furthermore, when function in the RTPJ is disrupted using a technique called transcranial magnetic stimulation (TMS), moral judgments reflect a reduced influence of mental states and a greater influence of outcomes: unintentional harms are judged as more forbidden, and failed attempts to harm are judged as more permissible (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, submitted). This pattern mirrors that observed in individuals with Asperger's Syndrome and five-year-old children, as described above. One source of developmental change in moral judgments may therefore be the maturation of specific brain regions for representing mental states such as beliefs. Consistent with this hypothesis is recent research suggesting the RTPJ may be late maturing (Saxe, Whitfield-Gabrieli, Scholz, and Pelphrey, submitted).

In all, this research provides reason to recognize the role of mental states in our explanations and evaluations of behavior, rather than just the role of stable character traits (whether they exist or not), the effects of people's actions on the external world, and the effects of the external world on people's actions.

### **The limits of empirical ethics**

Our results, among others, clearly suggest that adults do consider mental states when making moral judgments. Does the fact that we do so mean that we ought to? We think obviously not. As we argue below, science cannot provide positive support for our intuitions and theories about what ought to matter, morally. What science can do (and do well) is uncover problems. In the case of mental state based moral judgments, there are no social psychological experiments that undermine factual assumptions about beliefs and desires, as there are for stable traits. Nevertheless, there is another route by which scientific results could undermine faith in such judgments: by calling into question the underlying cognitive or neural processes. That is, some of our moral judgments might turn out to be psychological illusions.

(It's worth noting that this must be what Appiah means when he suggests repeatedly that science can correct our moral mistakes. Since Appiah never acknowledges that there may be a moral fact of the matter, what Appiah must mean by a 'moral mistake' is that we can be mistaken about *how* we make moral judgments. We might think that X is what matters to us, but science can tell us that our brains are truly responding to Y.)

Josh Greene claims to have found just this sort of evidence, concerning deontological (non-utilitarian) moral judgments (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Greene, Nystrom, Engell, Darley, & Cohen, 2004). In an fMRI study, Greene found that brain regions implicated in emotional processing showed greater activation for scenarios that typically elicited deontological judgments, i.e. that it is morally forbidden to harm one person in order to save many people. What made this experiment so sensational was that it seemed to show philosophers and ordinary people the limits of their introspective abilities when it came to the origins of their intuitions. Deontological intuitions are experienced as fundamentally moral judgments, but Greene claimed that these intuitions may reflect automatic emotional responses. The implication is that we should not endorse deontological intuitions if they are based on “emotion” rather than “reason” and specifically deontological reason; if there is no match between the underlying psychological process and the explicit output, then all bets are off.

Are moral judgments that distinguish between intentional and accidental harms problematic as well? Above, deontologists judge that pushing a man off a bridge so that his body stops a trolley from hitting five people (intended harm) is worse than turning a trolley away from five people and onto one person (foreseen harm). Consequentialists, however, claim to find it absurd that we judge intended harms as worse than foreseen harms. They claim that the justifications offered for these intuitions, in terms of mental states (e.g., intention), are merely post-hoc rationalizations. Any attempt to justify a moral difference between the two events is akin to arguing that the two lines in the famous Muller-Lyer illusion are of different length. While we may perceive one line to be

longer than the other, they are in fact the same length. Likewise, the pair of moral cases are the same: the agent knew exactly what he was doing, that is, harming one to save many. Intended and foreseen harms appear to be morally different only because of analogous flaws in our “moral sense”.

For judgments of intentional, accidental, and attempted harms, however, we think this analogy falls flat. The subtle mental state distinctions between intended and foreseen harms may be worth arguing over and indeed have fueled decades of philosophical debate (e.g., Kamm, 1998; Thomson, 1970). But the point that the agent’s mental state matters to moral judgment appears to be irresistible even to consequentialists. These consequentialists have argued that what makes intended and foreseen harms the same is that in both cases *the agent knew what he was doing* - that is, the agent’s basic mental state. It appears that we both have access to our mental state reasons for moral judgments and explicitly endorse these reasons.

Nor does the behavioral and neural evidence concerning the role of mental states in moral judgment point to a psychological illusion. The evidence points instead to a neat mapping of our neural responses and explicit explanations. The brain regions implicated in reasoning about mental states in a moral contexts are the very same regions implicated in reasoning about beliefs and desires in other contexts, like standard action-prediction false belief tasks described above (e.g., Saxe & Kanwisher, 2003; Saxe & Powell, 2006; Perner, Kronblicher, Staffen, & Ladurner, in press).

In sum, there is nothing in the evidence - from social psychology, cognitive psychology, or cognitive neuroscience - that undermines the commonsense moral intuition that the beliefs and desires matter for moral judgments. This is, of course, a

limited conclusion: there is nothing in the scientific evidence that *supports* this moral intuition, either. We think that this limitation reflects a general situation in the relationship between science and moral philosophy, to which we now turn for a few final thoughts.

### **The bottom line**

On the whole, the evidence above and ordinary intuition agree that the beliefs and desires matter for moral judgments (e.g., we forgive accidental harms and condemn failed attempts). That is, in general, healthy adults are both able to articulate mental state reasons for moral judgments, and also explicitly endorse the role of mental state reasons upon reflection. This - in and of itself - is interesting, as it could have turned out differently. The same is not the case for all kinds of moral judgments or all kinds of reasons for moral judgments; there are many instances where we cannot articulate the reason for our judgement, or where we actively disavow the reason once it's brought to our attention (Cushman, Young, & Hauser, 2006). Nevertheless, most observers find it hard to explain why mental states *ought* to matter. If Grace believed the white powder was sugar, why does that make her action (e.g., putting the powder in her friend's coffee) morally permissible? To say that she didn't intend harm is just to restate the premise. Why does it matter that she didn't intend harm? It just does.

This kind of response reflects a bottom line normative commitment. Many moral intuitions boil down to such a bottom line. For example, observers may assert that incest is wrong, even in the absence of possible procreation or psychological harm, but the same observers cannot explain why incest is wrong: "*It just is*". Jonathan Haidt has taken this kind of bottom line response to be evidence for 'moral dumbfounding' (Haidt,

2001). Confident as we may be in our moral intuitions, we are also often at a loss when we must explain or defend our attachment to them. Even Wittgenstein admits to this: “If I have exhausted the justifications, I have reached bedrock.... I am inclined to say: ‘This is simply what I do’” (Wittgenstein, 1953).

What does dumbfounding reveal about the normative status of intuitions?

According to Haidt, at least in some cases, dumbfound-able intuitions are suspect. If we cannot state the grounds for an intuition (because it arose out of an emotional bias, for example), then we should be less committed to that specific intuition. By contrast, we consider (and we think Appiah considers) dumbfounding a property of all moral intuitions and all moral theories at a fundamental level. For example, Appiah asserts that “the claim that we ought to do what’s in everyone’s long-term interest isn’t an evaluation, but a definition, a necessary truth that underlies morality” (pg. 24), a Humean bottom line. Analogously, the principle that we ought to aim for the greatest good for the greatest number appears to be the bottom line for utilitarians. Utilitarians can do no better job of explaining why the greatest good matters than ordinary observers can of explaining why incest is wrong.

If this view is right, then bottom line moral intuitions do not depend on facts, and so cannot receive support from facts or, a fortiori, from scientific facts. This does not render scientific facts irrelevant. As described above, scientific facts could still show that there is a false factual assumption, or a faulty cognitive mechanism, at work in the intuition. But if neither of these situations is true, then scientific results have nothing more to offer - as is the case, we argue, for our intuitions about beliefs and desires.

So what does science have to offer to moral philosophy? Science can undermine specific moral intuitions and moral theories, but its potential contribution ends there. Positive normative knowledge remains the domain of philosophy and, indeed, ordinary intuition.

### **Acknowledgments**

Thanks for Joe Paxton, Tamler Sommers, Josh Greene, Richard Holton, and Walter Sinnott-Armstrong for their helpful comments.

### **References**

- Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgment. In H. Guetzkow (ed.) *Groups, leadership and men*. Pittsburgh, PA: Carnegie Press.
- Casebeer, W. (2005). *Natural Ethical Facts: Evolution, Connectionism, and Moral Cognition*. Cambridge, MA: MIT Press.
- Cushman, F. (2008). Crime and Punishment: Distinguishing the roles of causal and intentional analysis in moral judgment. *Cognition*, 108, 353-380.
- Cushman, F., Young, L., & Hauser, M. D. (2006). The role of conscious reasoning and intuitions in moral judgment: testing three principles of harm. *Psychological Science*, 17(12), 1082-1089.
- Flavell, J.H. (1999). Cognitive Development: children's knowledge about the mind. *Annual Review of Psychology*, 50, 21-45.

- Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, 117(1), 21-38.
- Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44(2), 389-400.
- Greene J.D., Sommerville, R. B., Nystrom, L.E., Darley, J.M., Cohen, J.D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293, 2105-2108.
- Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review*, 108, 814-834.
- Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.
- Kamm, F. M. (1998). *Morality, Mortality: Death and Whom to Save From It*. New York: Oxford University Press .
- Kliemann, D., Young, L., Scholz, J., Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia*, 46, 2949-2957.
- Malle, B. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Cambridge, MA: MIT Press.
- Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal Psychology*, 67, 371-8.
- Piaget, J. (1965/1932). *The Moral Judgment of the Child*. New York: Free Press.
- Perner, J., Aichorn, M., Knronblicher, M., Staffen, W., & Ladurner, G. (2006). Thinking of mental and other representations: the roles of left and right temporo-parietal junction. *Social Neuroscience*, 1(3-4), 245-258.

- Ross, L. D. (1988). Situationist perspectives on the obedience experiments. *Contemporary Psychology, 33*, 101-104.
- Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *Neuroimage, 19*(4), 1835-1842.
- Saxe, R., & Powell, L. (2006). It's the thought that counts: Specific brain regions for one component of Theory of Mind. *Psychological Science, 17*(8), 692-699.
- Thomson, J. J. (1970). Individuating actions. *Journal of Philosophy, 68*, 774-781.
- Weiner, B. (1995). *Judgments of responsibility*. New York: Guilford.
- Wellman, H.M., Cross, D., Watson, J., 2001. Meta-analysis of theory-of-mind development: the truth about false belief. *Children Development, 72*, 655-684.
- Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Blackwell Publishing.
- Woolfolk, R., Doris, J., & Darley, J. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition, 100*, 283-301.
- Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences, 104*(20), 8235-8240.
- Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage, 40*, 1912-1920.
- Young, L., & Saxe, R. (in press). An fMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience*.
- Zimbardo, P. G. (1973). On the ethics of intervention in human psychological research: With special reference to the stanford prison experiment. *Cognition, 2*(2), 243-56.

