Phil 134/ 234 G
Moral Psych: W 18

Prof. Aaron Z. Zimmerman
March 5, 2018

Handout #4: Kennett and Fine on Self-Awareness, Self-Control and Moral Agency

1. Kennett's and Fine's (K&F's) Thesis

Self-Awareness and Self-Control are Necessary for Moral Agency: (1) K&F's Conceptual Thesis: "There can be no 'real' moral judgments in the absence of a capacity for reflective shaping and endorsement of moral judgments" (2008, 1). (2) K&F's Empirical Thesis: People fairly often do reflect on, reshape, and endorse those automatic moral judgments they sustain in the wake of self-conscious reflection, and so are moral agents who possess "real" moral judgments.

2. K&F's Interpretation of Haidt

According to K&F, Haidt's view entails that we are not moral agents:

(1) Experiments show widespread failures of self-knowledge: Our judgments about the morality or immorality of each other's actions are verbal expressions of emotions generated by processes to which we lack conscious access.
(2) Reflection on our concepts shows that self-knowledge is necessary for moral agency: To be moral agents properly praised and rewarded for our morally good deeds and properly blamed and punished for our bad deeds, we need to be have conscious access to the processes that give rise to our emotional reactions to the activities of ourselves and others—and not just the moral judgments we articulate on the basis of these emotional reactions—so that we can "reflectively shape" our moral intuitions and the cognitive framework they compose.
Therefore,
(3) Normative Skepticism: We are not properly praised for our good deeds nor properly blamed for our bad ones.

Questions: Is Haidt a normative skeptic? Does he reject the second premise in K&F's argument? If so, on what basis might he justify adopting a different attitude toward the behavior of young children (and non-human animals) when compared with the stance we take to supposedly "mature" adult humans? If self-awareness and self-control do not distinguish adults from children, why is it okay to imprison adults for harmful acts when similar treatment of a child or animal would be unwarranted?

"Small children and dogs and some non-human primates do have the capacity for empathy and yet they are not counted as moral agents. We suggest this is precisely because, and insofar as, they lack the developed rational capacities essential to such agency including the capacity to deliberate upon and regulate their automatic affective responses. Reasons responsiveness is central to morality in part because without it we cannot tell a satisfying story that makes sense of our practices of holding responsible. We systematically exclude individuals and categories of creatures from moral responsibility when we judge them to lack the capacity to respond to reasons as reasons and to guide their behaviour accordingly" (2007, 86)

Question: Are K&F right that Haidt's conception of moral judgment is incompatible with the conception of moral knowledge that underwrites or informs our normative practices of praise and blame?

3. Characterizing "Two-Systems" Views of Cognition

System 1: (a) System 1 processes information "automatically" or without much cognitive effort. (b) External stimuli (or features of the observable environment) exert the bulk of the causal influence on a System 1 process and its outputs. (c) System 1 processing often involves association. (d) The outputs of System 1 processes are intuitions or implicit attitudes. (e) When we are conscious of the outputs of System 1 processes it is normally through emotions or feelings associated with these outputs.

System 2: (a) System 2 processes information in a controlled, effortful fashion that requires cognitive effort. (b) The environment plays a relatively minor role in causally influencing System 2 processes and their outputs. (c) System 2 processing often involves deduction, induction, abduction, arithmetic computation and other forms of what we typically dub "reasoning." (d) The outputs of System 2 are explicit judgments or attitudes. (e) The outputs of System 2 are usually immediately available to consciousness and the processes can be accessed through introspection.

4. Tacit Cognition

We do not have direct first-person access to the overwhelmingly vast majority of our cognitive states. This class obviously includes the sub-personal states of highly modular cognitive faculties—for example the 2.5 D sketch which David Marr (1982) claims is crucially involved in the processing of visual stimuli. But it also seems to extend further to include more "informationally promiscuous" states of mind that are properly attributed to people (considered as whole organisms) and not just their cognitive parts. There are, for instance, various kinds of 'implicit memory' (Schacter, 1987). Long ago, Warrington and Weiskrantz (1968, 1982) showed that amnesiac patients retain memory traces of previously presented words despite an inability to explicitly recall having seen them. Wieskrantz (1986), in a now famous study, described patients with lesions of the striate cortex—so called 'blindsighters'—who sincerely report no conscious experience of relevant parts of their environment, but who perform better than chance when forced to guess about the shape and motion of stimuli in their visual fields. Prosopagnosic patients cannot explicitly recognize people they have met, but their emotional responses reveal memory-like representations with persisting cognitive effects (Bauer, 1984; Tranel and Damasio, 1985). Indeed, subsequent studies suggest that a great deal of person-level cognition may also be inaccessible (Marcel, 1983).

Moreover, unconscious cognitive states are not limited to cases of cognitive pathology. For instance, social psychologists Greenwald and Banaji (1995) obtained tentative evidence of widespread unconscious race-, sex- and age-based prejudice. A representative experiment asks subjects to depress the 'e' key when good words—such as 'happy' and 'friendly'—appear on their computer screen's center and to depress 'i' when bad words—such as 'miserable' and 'dangerous'—appear. Subjects are then asked to press one of these two keys when they see black faces on the screen's center and another when they see white faces. Various permutations of these tasks follow. In one variation subjects are asked to press 'e' when white faces or bad words appear and to press 'i' when black faces or good words appear. Another task asks subjects to press 'e' when white faces or good words appear and to press 'i' when black faces or bad words appear. Most white subjects find it difficult to group good words with black faces and bad words with white faces, but find it easy to group good words with white faces and bad words with black faces. That is, white subjects make more mistakes and/or take longer when trying to pair good with black and bad with white than when pairing bad with black and good with white. These results hold independently of the attitudes toward race these subjects avow on a questionnaire.

To test your own implicit attitudes on race and a variety of other issues you can go here:

https://implicit.harvard.edu/implicit/

Experiments of this kind are buttressed by reflection on ordinary cases of habituation, absent-mindedness and automaticity.

5. Controlling Implicit Prejudice

Question: What should you do when you have an implicit (System 1 generated attitude) that you explicitly disavow (because it conflicts with the results of your System 2 thought processes)?

K&F draw the following lessons from their survey of the prejudice control literature: Spontaneous judgments are more affected by implicit attitudes than judgments arrived at via conscious thought. Hurried ones are more affected than slow ones. Implicit attitudes reign when self-regulatory cognitive resources are either chronically low or diverted to other matters.

"Research also indicates that whether evaluations of others are based on activated stereotypes, or more accurate, individuating information, depends on motivation to be accurate, and availability of cognitive resources for controlled processing (see Kunda and Spencer 2003)."

Examples: Eye-gaze with regard to black people is better predicted by implicit attitude than self-report but preference for a roommate is better predicted by self-report than implicit attitude. Placement of chair with regard to a black person was better predicted by implicit attitude than self-report, but jury-like verdicts about responsibility for a crime were better predicted by self-report. (Dovidio et al. 1997; Rydell and McConnell 2006)

K&F criticize Haidt's social intuitionist model of moral judgment for employing an overly restrictive view of the role controlled processes can or do play in the generation and revision of our moral judgments (citing too Saltzstein and Kasachkoff 2004; and Fine 2006).

K&F organize their subsequent discussion of conscious control around Govorun and Payne's (2006, p. 130) distinction between **"after-the-fact correction"** and **"up-front mental control":**

A. After-the-Fact Correction

(a) "While a person's incidental mood or emotional state can 'contaminate' her moral judgments (e.g., Forgas and Moylan 1987; Wheatley and Haidt 2005) - something no rationalist would deny - this bias can be corrected more or less accurately (see Wilson and Brekke 1994), when the individual's attention is drawn to their mood as a possible source of bias (e.g., Schwartz and Clore 1983), or she is motivated to be accurate (e.g., Lerner et al., 1998; although see Payne et al. 2005a for evidence that bias correction cannot always be successfully achieved).

(b) "Gabriel et al. (2007) assessed automatic (using the IAT), cognitive and affective attitudes towards homosexuals in a sample of students. The cognitive attitude scale tapped beliefs about homosexuality (for example, that female homosexuality is a sickness, or that gay men should not be allowed to work with children). The affective attitude scale, by contrast, tapped emotional responses to homosexuality, including experiences that Haidt might term moral disgust or other moral emotions. In this questionnaire, the participant indicates how much discomfort he would feel if, for example, he learned that his son's teacher was gay, or if he saw two lesbians kissing. For participants who indicated only low internal motivation to control prejudiced responses,

automatic attitudes correlated with both cognitive and affective attitudes, as the SIM would predict. But for participants with a high internal motivation to control prejudice, increasingly negative automatic attitudes towards homosexuality manifested neither in more negative affective attitudes nor more negative cognitive attitudes. Participants were then given the opportunity to sign a petition to maintain funds and to donate money to a (real) local gay organization facing the prospect of having public funding discontinued. Contrary (we assume) to the predictions of the SIM, cognitive attitudes, but not affective attitudes, predicted helping behavior" (2007, 90).

Haidt's response: "The tight connection between flashes of intuition and conscious moral judgments ... is not inevitable: Often a person has a flash of negative feeling, for example, towards stigmatized groups ... yet because of one's other values, one resists or blocks the normal tendency to progress from intuition to consciously endorsed judgment" (Haidt and Bjorkland 2007a, p.818).

B. Up-Front Pre-Conscious Control

(a) "Keith Payne uses a paradigm in which a sequence of guns and tools appear on the computer screen. The volunteer's task is to categorize the object with a key press, as quickly as possible. Before each trial a male face appears that is either black or white. Both automatic and controlled processes contribute to responses on the weapon-identification task, and the experimental design enables separation of automatic and controlled contributions to performance. For most participants, an automatic 'black man-danger' association will, in the absence of control, result in 'false positive' errors on trials in which a black face is followed by a harmless tool. However, Payne (2005) has found that people with a high capacity for self-regulatory control, and stronger motivation to control prejudice, show less behavioural expression of automatically activated associations; that is, they show fewer 'false gun' errors after seeing black faces. Importantly, this is not because negative automatic attitudes are any less strongly activated in these individuals. Rather, they exert greater cognitive control which, according to Payne (2005, p. 491), enables people to 'constrain their processing to the relevant information rather than being driven by irrelevant but activated information.'"

(b) Amodio et al. (2008), using Payne's weapon-identification task, have collected electroencephalographic (EEG) data suggesting that the conflict between an activated stereotype and the goal to avoid stereotyping is detected preconsciously. Amodio et al. found that a burst of 'conflict monitoring' neural activity takes place about 100ms before volunteers successfully avoid making a mistaken call of 'gun' following a black face. Furthermore, people who were motivated to control prejudice because of their own internal moral standards showed larger waves of conflict-monitoring EEG activity, and fewer false positive errors, than volunteers whose prejudice-control motivations stemmed more from concern about how they came across to others. Amodio et al. (2008, p. 72) suggest that '[t]hese findings show that effective response control may be deployed without a person's awareness that a race-biased response was averted.'

(c) "Stewart and Payne (2008) asked one group of volunteers to make the counter-stereotypical commitment that, 'Whenever I see a Black face on the screen, I will think the word, 'safe'.' These volunteers made fewer false positive gun responses following a black face, compared with groups who made non counter-stereotypical commitments. Further analysis of the data revealed that, unlike the findings of Payne's previous work, this was not due to increased cognitive control. Rather, it arose from a reduction in biased automatic activations. Stewart and Payne (2008, p. 1344) note that this conscious commitment strategy provides 'one way in which conscious strategies can be used to overcome automatic stereotyping even when all the right circumstances (e.g., opportunity for controlled thinking, awareness, etc.) are not in place'" (2007, 93).

Conclusion: "While we concur with Haidt that many of our moral judgments will be based on intuitive responses, we argue that in certain situations, in certain individuals, those very same moral judgments will not be based on automatic evaluations of events or people. We may effortfully over-ride judgments based on moral intuitions, discount moral emotions that we believe to be irrelevant or misplaced, and exert preconscious control such that the activated associations of our moral intuitions do not interfere with the processing of more relevant information" (ibid.)

6. K&J's Rationalistic Conception of Moral Judgment

K&F report James Rachel's conception of moral judgment with approval. (They attribute a similar conception to John Deigh, Michael Smith and R. J. Wallace.)

A Rationalist Constraint on Moral Judgment: If S **does not** have reasons for thinking that x is wrong, then S's judgment that X is wrong is not a moral judgment.

Question: Is this a correct view? If someone is raised to believe that (e.g.) masturbation is wrong, mightn't he or she genuinely believe that this activity is immoral without being able to supply a reason to back up her view?

According to K&F, "When an individual makes a moral judgment, it is plausible to suppose that the reasons implied or adduced in support of the judgment must be reasons which the agent herself (albeit perhaps mistakenly) takes to justify and not merely to explain the judgment. Otherwise the judgment can have no normative authority for her. On the face of it, judgments determined by automatic processes that occur below the level of consciousness would seem to lack this authority and so could not count as the agent's 'real' moral judgments, at least absent a process of reflective endorsement through which the agent identifies and aligns herself with the considerations that she takes to justify the initial automatic moral attitude" (81).

Question: Why do K&F feel pressed to say that judgments an agent cannot defend are not "real"? Isn't it more "intuitive" to say that beliefs or judgments of this kind are **real** but **unjustified**?

K&F's response: "We think that the normativity of moral judgment is most plausibly cashed out in terms of reflective endorsement and regulation so our claim is that in cases of conflict the agent's considered view deserves the title of the 'real' or authoritative moral judgment even when we think their automatic responses better track the reasons that there are. But as we have suggested above this does not require that our each and every moral response must be the product or subject of explicit effortful deliberation in order to so count. Plainly we do not have the cognitive resources to devote to such a task. Nevertheless, we think our reflective views of 'the good, defensible and fulfilling life' are capable of regulating our moral responses, directly or indirectly, so that spontaneous intuitions that do not accord with our reflectively endorsed evaluative framework may be modified, overridden or set aside" (86).

K&F's interpretation of Haidt on this matter: "Haidt's position would appear to be, then, that to know a person's (current) moral intuition about a person or situation is to know virtually all that we need to know about their moral stance at that time. To ask them to reason about their position is merely to assess their facility either at confabulating reasons to support their intuitions, or their ability to trigger new moral intuitions in others" (2007, 83-4).

A Skeptical Constraint on Moral Judgment: If S **does** have reasons for thinking that x is wrong,

then S's judgment that X is wrong is not a moral judgment.

Questions: Is Haidt really a skeptic about moral judgment in the way described above? Isn't the more reasonable view that some but not all of the average person's moral judgments will have the backing of reasons?

7. Reason Tracking v. Reason Responding

A distinction drawn by Karen Jones (2003).

S **tracks the reasons** to believe that p (i.e. good evidence for p) just in case S regularly believes p when she has good evidence that p and reliably fails to believe p when she doesn't.

Note: Reason-tracking is exhibited by non-human animals when they (e.g.) reliably come to believe that there is food nearby when they smell food etc. It does not require System 2 mechanisms.

S **responds to the reasons** to believe that p just in case S regularly believes that she has good reasons to believe that p when she does and comes to believe that p on that basis and S regularly believes that she lacks good reasons to believe that p when she lacks such reasons and refrains from believing that p on this basis.

Note: Reason-responding would seem to require the concept associated with the word "reason" (or "evidence") and so it isn't likely that an animal can respond to reasons without employing System 2.

According to K&F, "An implication of Haidt's work is that our first personal experience of ourselves as reason responders is illusory" (2007, 85)

Question: Is this fair to Haidt? Does he deny the existence of evidence for or against adopting a given moral view? Does he argue that explicit regard for reasons or evidence does not play enough of a role in the processes through which we adopting and revise our moral views for us to be credited with agency?

Jones's Thesis: Sometimes the reasons you merely track are better evidence than the reasons to which you respond. For example, someone might seem creepy even though you don't have articulate or consciously accessible evidence that he is dangerous or bad in some other way.

Kahneman's example: an experienced nurse or doctor can sometimes accurately diagnosis cardiac difficulty in a patient from the nurse's or doctor's intuitive sense of the patients health even when more objective tests prove inconclusive. K&F also discuss the fictional case of Huck Finn who helps his friend Jim escape slavery even though he believes he is doing something wrong (in allowing Jim to break the law and rob his master of her property).

A Difficult Question: Perhaps your gut is sometimes more reliable than your conscious judgment. But (as cases of implicit prejudice make clear) sometimes your conscious judgment is more reliable than your gut. Given this, how should a person decide, in any given case, whether she should trust her gut or her more deliberate reasoning (i.e. her system 1 or her system 2)?

8. Psychopathology and Autism

According to K&F, **psychopaths** lack both proper affective response (moral reason-tracking capacities) and proper self-awareness and self-control (moral reason-responding capacities).

Those diagnosed as strongly but fairly **high-functioning autistics** lack robust perspective taking and normal affective responses (moral reason-tracking capacities) but can develop the proper self-awareness and self-control necessary to be guided by moral reasons.

They argue on this basis that autistics can be moral agents but that it is unclear whether psychopaths can be.

Questions: Is this the correct stance to take toward these populations?