Jonathan Haidt and the Disappearance of Moral Knowledge:

How Good Intentions and Philosophical Confusions* Threaten to Perpetuate the Problem.

Aaron Preston
Biola University

Prepared for the Symposium on Dallas Willard's
The Disappearance of Moral Knowledge
Biola University's Center for Christian Thought
May 14-15, 2019

Beginning with the 2006 publication of his *The Happiness Hypothesis*, Jonathan Haidt has emerged as a leading public intellectual addressing matters of morality and ethics, and doing so in a way that is particularly relevant to the disappearance of moral knowledge. A social psychologist by training, Haidt specializes in moral and political psychology, and has brought his expertise to bear upon a broad range of issues concerning the relationship of morality to human flourishing and happiness on both the individual and social levels.

It is precisely because of Haidt's prominence as a public intellectual that his work merits attention from those of us concerned about the disappearance, and possible recovery, of moral knowledge as a publicly available resource. Within the academic field of moral psychology, there may be better thinkers with better theories. But none of them have broadcast their views to the general public as successfully as Haidt. As far as the project of making moral knowledge available as a public resource is concerned, Haidt is the one who is making an impact.

And for the most part, I think his impact has been good. I am especially grateful for his second and third popular books, which are aimed at restoring some sanity and civility to public discourse and institutional life in the United States. His second book, *The Righteous Mind* (2012), sought to address the problem of political polarization by generating mutual understanding between liberals and conservatives. One's political leanings, Haidt argues, are in part a product of one's personality, which is in part a product of one's biology. We therefore shouldn't demonize each other for our political preferences, any more than we would blame someone for their basic personality traits. Self-righteousness and tribalism are also natural inclinations, but ones we should do our best to counteract given that they are counterproductive vis-à-vis the aim of maintaining a stable and peaceful common life.

Instead of taking sides and squaring off against each other, we should try to appreciate that liberals and conservatives have different ways of caring about a range of almost universally recognized goods. Haidt argues that most peoples around the world have been found to care

about goods or values in six different categories, all of which (he claims) have an evolutionary basis:

1. care/harm
2. liberty/oppression
3. fairness/cheating
4. loyalty/betrayal
5. authority/subversion
6. sanctity/degradation

Liberals in the contemporary West tend to focus almost exclusively on the first two, whereas (classical) conservatives characteristically have a more balanced appreciation for the full range of moral goods. (211) While Haidt thinks the liberal's moral "tunnel vision" has served a useful purpose in the past by giving special support to generally undervalued goods, Haidt recognizes that from the standpoint of cross-cultural moral psychology one might reasonably deem it a deficient moral perspective. One of the most controversial moves in the book is Haidt's attempt to alert liberals to the fact that they are idiosyncratic and indeed WEIRD: Western, educated, industrialized, rich, and democratic. Most of the world is not like this, and is in fact more like Western conservatives in their moral sensibilities. Thus, rather than demonizing conservatives, Haidt counsels liberals to recognize the deeply human wisdom embedded in the conservative position even while they retain their special commitment to care and liberty. In a remarkable instance of personal transformation which attests to Haidt's intellectual integrity and moral seriousness, Haidt has taken his own advice by abandoning the progressive liberalism of his youth and becoming a political centrist and an advocate for "viewpoint diversity".

These same findings about the full scope of ordinary human moral concern have led Haidt to worry about something like the disappearance of moral knowledge. In chapter 8 of *The Happiness Hypothesis* we see Haidt worrying over the "turn from character ethics to quandary ethics" (164) in the history of ethical theory. Ancient approaches to ethics, he explains focus on the cultivation of virtues, "rely heavily on maxims and role models rather than proofs and logic," (159) and "emphasize practice and habit rather than factual knowledge." (160)

This may sound as if Haidt is being dismissive about the role of factual knowledge in the moral life. But in truth he only means to emphasize the importance of habituation to the moral life, which he understands in terms of encoding "tacit knowledge" into our automatic response systems. This, of course, is exactly the point of Willard's teaching on the spiritual disciplines. Both he and Haidt recognize that the modern mind has lost touch with ancient wisdom about the need to shape the largely unconscious structures of the self, out of which many of our actions and reactions spontaneously emerge, and both have embraced the task of trying to reintroduce this wisdom to different audiences. Haidt does this by developing the metaphor of the self as a rider on an elephant:

> modern theories about rational choice and information processing don't adequately explain weakness of the will. The older metaphors about controlling animals work beautifully. The image I came up with for myself… was that I was a rider on the back of

> an elephant. I'm holding the reins in my hands, and pulling one way or the other I can tell the elephant to turn, to stop, or to go. I can direct things, but only when the elephant doesn't have desires of his own. When the elephant really wants to do something, I'm no match for him. (HH, 4)

The most serious problems with Haidt's perspective are grounded in the way he works out the details of this metaphor, and we shall turn to them further on. But here we simply note that Haidt and Willard agree that there is more to morality than factual knowledge and discursive moral reasoning. As Haidt explains immediately following his observation that ancient theories "emphasize practice and habit rather than factual knowledge,"

> "if moral instruction imparts only explicit knowledge… it will have no effect on the elephant [i.e., the part of ourselves that acts automatically, from a place of habit or instinct]. Moral education must also impart tacit knowledge – skills of social perception and social emotion so finely tuned that one automatically feels the right thing in each situation, knows the right thing to do, and then wants to do it. Morality for the ancients was a kind of practical wisdom." (160)

By contrast, modern moral philosophers have characteristically focused on action rather than character, have agreed that "moral decision-making requires logical reason and sometimes even mathematical calculation," (163) and have therefore sought "a single moral rule, to be applied through the power of reason, that can separate good from bad" (161). The rational testing of proposed rules in turn has led to what Edmund Pincoffs calls "quandary ethics" - undue emphasis on solving uncommon and usually unrealistic puzzle-cases and ethical dilemmas, rather than focusing on how to live in the conditions of ordinary life. Haidt's rejection of this approach is reminiscent of Willard's discussion, in DMK chapter 8, of Western philosophy's "persistent tendency to try to force moral knowledge into a model or form of knowledge which it simply cannot assume," specifically, the form of a deductive science: "one logically organized totality… *logical* unity under *theory*… a "deductive" order in some fairly rigorous sense." (DMK, 356)

Like Willard, Haidt sees the turn "from character to quandary" as a "profound mistake." (HH, 164) it gives rise to a "thin and restricted" conception of morality concerning "trade-offs between self-interest and the interests of others," (164) and it has no natural place for moral training in the form of habituation. Like Willard, Haidt sees this as the "great omission" from contemporary approaches to moral education.

Between the modernist restriction of moral education to "moral reasoning" and the liberal restriction of "morality" to matters of care/harm and liberty/oppression, Haidt observes that "we have indeed lost something important – a richly textured common ethos with widely shared virtues and values." Life in a virtue-oriented culture is "a dense web of moral fibers." (175) but in the modern world where ethical theory is focused on quandaries, and ordinary life is left to find its own way, we run into Durkheim's problem of anomie:

> Anomie is the condition of a society in which there are no clear rules, norms, or standards of value. In an anomic society, people can do as they please; but without any clear

> standards or respected social institutions to enforce the standards it is harder for people to find things they want to do. Anomie breeds feelings of rootlessness and anxiety and leads to an increase in a moral and antisocial behavior. (177)

Like Willard, and even more strongly, Haidt acknowledges that there have been up-sides to the overthrow of "traditional morality": society has in many ways become more inclusive and humane. But still, something important has been lost. In recognizing this, I believe Haidt is worrying over essentially the same condition that Willard labels "the disappearance of moral knowledge".

What does Haidt propose we should do about this? Does he propose a return to classical, virtue-oriented approaches to morality? Not quite. *The Happiness Hypothesis* includes a chapter on virtue (ch. 8), and he does think that the pursuit of virtue in the manner of contemporary positive psychology is an important ingredient in the recipe for human happiness. But it is happiness itself, or more broadly *emotion*, rather than virtue, that is the center of gravity in Haidt's moral world.

This is why Haidt selects Hume as his philosophical hero, and takes as his motto Hume's statement that "reason is and ought only to be the slave of the passions." This is one of several instances in which Haidt chooses the wrong philosophical ally, and he does so because of confusions about the natures reason, intuition, and emotion. These confusions infect his readings of Hume and Plato, as well as his own attempts at clarifying his rider-elephant metaphor. In the remainder of this paper, I will focus mainly on the latter issue,

The fundamental distinction Haidt means to capture with this metaphor is the one central to dual processing models of cognition: the distinction between consciously controlled processes and automatic processes in the human self. Haidt explains: "most mental processes happen automatically, without the need for conscious attention or control." In fact, "Most automatic processes are completely unconscious, although some of them show a part of themselves to consciousness…." By contrast, controlled processes are "the kind of thinking that takes some effort, that proceeds in steps and that always plays out on the center stage of consciousness." (HH, 14-15)

So far so good. However, problems begin to emerge when Haidt observes that "moral judgment is like aesthetic judgment":

> When you see a painting, you usually know instantly and automatically whether you like it. If someone asks you to explain your judgment, you confabulate. You don't really know why you think something is beautiful, but [the rider] is skilled at making up reasons…. You search for a plausible reason for liking the painting, and you latch onto the first reason that makes sense….

And then:

> Moral arguments are much the same: two people feel strongly about an issue, their feelings come first, and their reasons are invented on the fly, to throw at each other.… If you listen closely to moral arguments, you can sometimes hear something surprising: that

> it is really the elephant holding the reins, guiding the rider. It is the elephant who decides what is good or bad, beautiful or ugly. Gut feelings, intuitions, and snap judgments happen constantly and automatically…, But only the rider can string sentences together and create arguments to give to other people. In moral arguments, the rider goes beyond being just an advisor to the elephant; he becomes a lawyer, fighting in the court of public opinion to persuade others of the elephant's point of view. (HH, 21-22)

Two key problems are on display in this passage. First, Haidt seems to be saying that reasoning always and only takes the form of *post hoc* rationalization. Rather than a guide to truth, reason becomes a shield against the truth. Second, note how Haidt shifts from describing moral judgments as automatic or spontaneous processes to describing them as "feelings." He also describes them as "intuitions," but in ordinary parlance nowadays "intuition" just means "gut feeling." Listing the two-terms side by side suggests that they are basically two ways of saying the same thing, as ordinary parlance has it. This is reinforced by assimilating moral judgment to aesthetic judgment, which can be counted on to suggest to many readers that the former are like the latter not merely in being automatic rather than controlled, but also in pertaining to matters of subjective taste, not susceptible to rational evaluation and correction.

The problem with Haidt's characterization of moral judgments as "feelings" is that, understood in accordance with ordinary parlance, this makes morality look a lot like what emotivism in ethical theory says it is, as discussed in Chapter 4 of DMK. Emotivism is the view which holds that moral claims have no cognitive content, and are therefore susceptible neither to rational understanding nor rational evaluation. They serve only to express feelings and make emotional appeals to others. In fact this is not Haidt's view, but presumably in an effort to keep things simple for nonacademic readers, he has sent a false signal which, by driving readers in the direction of emotivism, threatens to perpetuate the disappearance of moral knowledge.

In his academic writings, Haidt is much more careful. In an early article[1] co-authored with Richard Schweder (then his mentor on a post-doc at the University of Chicago), Haidt distinguishes among cognitivism, emotivism, rationalism, and intuitionism as these terms/views are understood among psychologists:

> EMOTIVISM
>
> The view that "the moral qualities of events in the world do not exist aside from people's sentimental reactions to those events, including their feelings of approval and disapproval. For emotivists, morality (like beauty) is in the mind of the arbiter, moral appraisals are not subject to determinations of truth and falsehood, and they cannot be judged against rational standards. They are simply declarations of preferences and values." (361)

[1] Schweder, R. and Haidt, J. 1993. The future of moral psychology: truth, intuition, and the pluralist way. Psychological science. 4:6, November 1993.

COGNITIVISM

> "Moral cognitivism is the position that qualities such as goodness, rightness, justice, or beneficence are real and knowable and that moral statements can therefore be either true or false…. A cognitivist approach to moral development tries to identify the particular mix of intellectual skills… and interpersonal experiences… that makes it possible to apprehend or figure out moral truth. The basic point of cognitivist theories of moral psychology is that everyday moral appraisals… can be right or wrong; they are not subjective or inculcated tastes, opinions, or attitudes, as emotivism would have it." (360)

So far so good – this is pretty much how philosophers understand these terms as well. Not so with the way they draw the distinction between two species of Cognitivism: Rationalism and Intuitionism.

RATIONALISM

> "theories of cognitive rationalism argue that knowledge of moral truth comes from a process of argumentation and deductive reasoning." (363)

INTUITIONISM

> "Theories of cognitive intuitionism assume that moral properties… are objective and knowable but apprehended as self-evident truths. … Such theories assume that … objective moral qualities are neither logical properties nor observable properties [but] objective properties of a different kind, properties that are open to the rational intuitive capacities of the human mind or nervous system. They can be activated without reflection. They are not dependent on deliberative reason or argumentation to bring them online." (363)

In philosophy, the historically standard view is that intuition is a rational power – as Haidt and Schweder seem to affirm in speaking of "rational intuitive powers." But for philosophers, this precludes any opposition between "rationalism" and "intuitionism." Both are species of rationalism. On the historically standard view, humans have both discursive reason and intuitive reason, where the letter is (as the etymology of the term suggests) the power of direct insight into truth. What's more, for those who recognize both powers in human beings, they are usually taken to work together, and the intuitive reason is usually taken to be the ultimate source of knowledge – it is what enables us to grasp first principles, including the axioms and rules of inference by which discursive reasoning proceeds. Consequently, to oppose intuitionism to rationalism appears bizarre from this perspective.

Even more bizarre is Haidt's tendency in his popular writings to cast Plato as a "rationalist" in his discursive-reasoning-limited sense and on that basis to reject his approach to morality, while aligning himself with Hume and making him the poster boy for the "social intuitionist model" of moral psychology (and moral knowledge). Hume was an *emotivist*, pure and simple. For him, emotions are not cognitive states, but raw qualitative "feels" (what we now call *qualia*) with no representational content or function, and there is no "interface" for reason to interact with emotion. It is precisely this perspective that informed 20th C. emotivism's understanding of

emotions as devoid of cognitive content, the blind drives behind our meaningless moral utterances, and which helped to propel the meaning of "intuition" in the direction of "gut feelings."

In the 1993 paper with Schweder, Haidt seems to be aware of this. They speak of "the emotivist legacy of David Hume" and note that Hume presents us with a false dichotomy in supposing that the grounds of morality must be found either in reason (i.e. rationalism) or in sentiment (i.e., emotivism). (361) This is a false dichotomy because "one can develop a cognitivist theory … about any kind of mental state", including emotions:

> "There are cognitivist theories of the emotions and emotivist theories of inductive reasoning. The mark of a cognitivist theory is the assumption that mental states serve primarily a representational function. The mark of an emotivist theory is the assumption that mental states serve primarily nonrepresentational functions." (361)

But if representation is the mark of the cognitive, then it's fair to say that neither thoughts nor feelings are cognitive for Hume, since for him thoughts and feelings (ideas and impressions) are treated as the objects of consciousness, and are not themselves taken to be representations of other things. Even though Hume is often labeled a "representationalist" because he thinks that the objects of thought are contents of the mind rather than external objects, the fact is that, because he lacks any account of intentionality as a property of impressions and ideas, mental contents for him are not properly "representations". So Hume cannot be a cognitivist about any metal states, let alone emotions.

By contrast, in his academic work, Haidt consistently identifies himself as a cognitivist intuitionist who embraces a cognitive theory of the emotions. Emotions for him – but not for Hume – are automatically-occurring representational states, which makes them intuitions; and they are "affect-laden" intuitions, which make them emotions. That is, emotions are, for Haidt, a subclass of cognitive intuitions distinguished from other intuitions by their inclusion of a salient affective component.

That's all fine and good' but in his popular work Haidt ignores this extremely significant difference between Hume's views and his own, persistently holding up Hume as his hero and using as his motto Hume's dictum about reason being the slave of the passions. In *The Righteous Mind*, Haidt devotes a paltry 3 pages (52-54) to explaining his view of emotion-as-cognitive-intuition, but this is just is not enough. As he and Schweder put it in the penultimate line of their 1993 essay: "The mind of the moralist is located in the heart, which is, *paradoxically*, a cognitive organ, and it is through the heart that we come to know moral truths." (364, my emphasis) Why "paradoxically"? Because that is not normally how we think about the heart, the emotions. We normally think of them in noncognitive terms. And it will take a lot of doing to get people to change their mental habits on this. A short three-page discussion will not do it.

At this point a brief aside is in order, as this point is relevant to the main criticism raised against DMK in what is, to my knowledge, the only academic review the book as yet received. In Notre Dame Philosophical Reviews, Adam Pelser criticizes "Willard's apparent uneasiness about

emotions … preferring instead to" use the terms "attitudes" "intuitions." "This uneasiness about emotions… hampers Willard's moral epistemology," Pelser says, and "emotion" would have been a better choice, because "although emotion is often misunderstood, at least it is not a mysterious faculty like "intuition."" But actually, intuition is only mysterious for those whose conceptual horizons are limited by the analytic tradition in philosophy. For realist phenomenologists like Willard, there's nothing particularly mysterious about intuition. Failing to recognize this, Pelser continues:

> Perhaps Willard is wary of identifying emotion by name as an important part of his moral epistemology because of the way its name has been tarnished by the Emotivists and other Noncognitivists.

That is exactly correct. What's interesting is that Pelser acknowledges that this has created a serious difficulty, that it has in fact given "emotion" a bad name. But, unlike "intuition" he thinks that "emotion" is a name worth rehabilitating:

> I submit that we will not be able to make the kind of progress Willard desires toward developing an adequate account of moral knowledge and articulating a compelling vision of the moral life unless we engage in the difficult moral-psychological project of understanding emotion and restoring emotion's good name. A robust epistemology of emotion thus would go a long way toward strengthening and filling out Willard's account of moral knowledge

Given that "emotion" is badly misunderstood on account of the influence of Emotivism and Noncognitivism, and that the "project of understanding emotion and restoring emotion's good name" is expected to be "difficult", is not clear to me what advantage "emotion" has over "intuition" vis-à-vis the project of "developing an adequate account of moral knowledge and articulating a compelling vision of the moral life." Both are terms that have had their meanings decisively skewed by emotivism; but only "emotion" contains the basis for that skewing within itself, since, even if emotions are cognitive, they are still essentially affect-laden in a way that intuitions in the cognitive sense are not. And the fact that "intuition" literally means "insight" gives us a leg to stand on in insisting that it is a mistake to identify intuitions with gut feelings. My own view is that "emotion" is at this point in history probably not redeemable for purposes of communicating a sound moral epistemology to a popular audience, and what's needed is to get people to step outside the circle of what they regard as their "emotions" to evaluate them from what they would regard as a detached, unemotional standpoint.

In any case, Haidt seems to have jumped into this mess with both feet since he wants to use both concepts – emotion and intuition – and to say that the one is a species of the other. Given the damage done to both concepts under the dominion of emotivism in the mid 20th century, It will take more than a short three-page discussion to get the right ideas across to a general reader.

Another problem with *The Righteous Mind*'s short discussion of intuitionism in is that it links cognition not to representation, as Haidt and Schweder had done in their 1993 article, but to information processing: "cognition just refers to information processing… [and] emotions are kind of information processing," therefore emotions are cognitions (52, 53). Not only is this

false (information processing is not cognition, as Searle's Chinese room thought experiment shows[2]), it doesn't even begin to illuminate the possibility that emotions are cognitive in a way that might open them to rational scrutiny - which must be done if we are to rehabilitate "emotion" from its corruption at the hands of the emotivists. Compare Haidt's statement to Martha Nussbaum's description of emotions as "intelligent responses to the perception of value," which "include in their content judgments that can be true or false, and good or bad guides to ethical choice" (from *Upheavals of Thought*).[3] This description of emotions as including "in their content judgments that can be true or false," helps to dispel the haze of *emotivism* that distorts the popular understanding - or misunderstanding - of what emotions are. Haidt's description simply does not do this. Nor does anything else in his all-too-short discussion.

So, again, my concern is that by continually speaking of "emotion" as the central category for human moral psychology without really bending over backwards to reclaim the concept of emotion from the emotivists, and by repeatedly linking his views to Hume, whose views are the headwaters of 20th century emotivism, Haidt sends a false signal which, by driving readers in the direction of emotivism, threatens to perpetuate the disappearance of moral knowledge.

What is needed, I think, is greater attention to the phenomenology of emotion and intuition. If we are to reclaim either concept from the emotivist paradigm that now dominates ordinary thought about emotion and intution, we have to give people rich descriptions of these mental phenomena that enable them to recognize the cognitive natures or dimensions of intuition and emotion from their own, introspective experiences. This is also what's needed to answer the objection to DMK's reliance on "intuition" rather than "emotion" in developing its own moral epistemology. If this work could be done adequately, it seems to me that Haidt's approach to moral knowledge and Willard's can be taken as complementary in a number of ways: both are fundamentally cognitive intuitionist views that confer a prominent role upon "attitudes" or "emotions" in establishing a basis for moral knowledge. Both aim to expand moral knowledge by gathering data about how people see the world in moral terms - although Willard is focused more narrowly on how people identify moral exemplars and on how moral exemplars see their world and those around them, whereas Haidt is interested in a broader range of moral phenomena. It strikes me that Willard's proposal for building a science of ethics fits nicely as an extension of positives psychology's cross-cultural research program on universal human virtues, which Haidt endorses. And so on. Given Haidt's prominence in his field and a public intellectual, perhaps the closest thing to we have to a public expert on moral knowledge, he

---

[2] For a more careful and sophisticated argument to the same effect, see Horst, Symbols and Computation: A Critique of the Computational Theory of Mind, *Minds and Machines* **9:** 347–381, 1999.

[3] Now, I don't agree with this characterization of the emotions – although it's much closer to the truth than anything Haidt says about them. I think judgments are not among the *contents* of the emotions themselves, but that they are necessarily connected external concomitants of the emotions. Just as shape and size are necessarily connected but neither is a "content" of the other, evaluative judgment and emotion are necessarily connected but distinct in their essences. The relation between them is *presupposition*: emotions *presuppose* value judgments. We can even say that the judgments are implicit in the emotions – but all that means is that the presence of the emotion is inexplicable apart from the judgment. To make the judgment explicit is to step outside the emotion itself. All of this, it seems to me, is what you get when you attend carefully to the phenomenology of emotion. But of course there is room for disagreement here.

would make a good ally for the Willardian project… but only if Haidt can get his fundamental confusions about reason, intuition and emotion sorted out.