Journal of
Cognition
and
Neuroethics

The Theory-Theory of Moral Concepts

John Jung Park
Christopher Newport University

Biography

My interests lie in moral psychology, metaethics, and philosophy of mind. I am particularly interested in the structure of our moral concepts and what implications this has for metaethics. I am also interested in the hard problem of consciousness.

Acknowledgements

I would like to thank David Wong, Karen Neander, Daniel Weiskopf, and Wayne Norman for their input on this paper.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). September, 2015. Volume 3, Issue 1.

Citation

Park, John Jung. 2015. “The Theory-Theory of Moral Concepts.” *Journal of Cognition and Neuroethics* 3 (1): 117–138.

The Theory-Theory of Moral Concepts

John Jung Park

Abstract

There are many views about the structure of concepts, a plausible one of which is the theory-theory. Though this view is plausible for concrete concepts, it is unclear that it would work for abstract concepts, and then for moral concepts. The goal of this paper is to provide a plausible theory-theory account for moral concepts and show that it is supported by results in the moral psychology literature. Such studies in moral psychology do not explicitly contend for the theory-theory of moral concepts, but I demonstrate that they actually do provide evidence for the use of theory knowledge at times in moral categorization and decision-making. In philosophy of cognitive science, I newly show that there is evidence that the theory-theory does apply to some moral concepts.

Keywords

Moral Concepts, Concepts, Mental Representations, Theory-Theory, Prototype Theory, Moral Cognition, Empirical Moral Psychology, Cognitive Science

Introduction

The theory view for concrete concepts claims that concrete concepts are mental representations of hidden essences, causal laws, functions, explanatory relations, and/or general background knowledge (Carey 1985 and 2009; Murphy & Medin 1985; Keil 1989; Gopnik & Meltzoff 1997).[1] As my principle aims in this paper are to freshly explicate the theory view for moral concepts and provide evidence for it in light of moral categorization, I perceive my contribution in this case to be primarily in the concepts literature, where theorists in part address what constitutes concepts (Rosch & Mervis 1975; Prinz 2002; Machery 2009; Weiskopf 2009), rather than in the causal moral judgment literature in moral psychology, where moral psychologists examine what mental states influence the making of moral judgments (Greene et al. 2009; 2013; Mendez 2005; Cushman 2008; Young & Saxe 2008; Haidt 2012). The first reason for this is that the causal judgment literature rarely explicitly discusses the structure of moral concepts. Second, some of the main aims of this paper are to elaborate upon the theory view in the concepts literature and show how such a view applies at times to the

[1] Mental representations are mental states that refer to or purport to represent things in the world. For example, my concept DOG refers to the category *dog*.

moral concepts domain. Hence, I see my contribution that establishes the viability of the theory-theory for moral concepts as being primarily in the concepts field. To note, I do not claim that moral concepts only have theory structure. I perfectly leave open the possibility that moral concepts can store many other different kinds of knowledge that can be used individually or conjointly in cognition. However, due to obvious space concerns, I will focus exclusively on the issue of whether some moral concepts may have theory structure.

Psychological concepts are generally understood as being the constituents of thought or as Locke states, they are the "materials of reason and knowledge." They are the basic units of the human understanding. For example, my judgment RAPE IS WRONG is made up of three individual concepts: RAPE, IS, and WRONG.[2] Also, concepts are understood as being mental representations or bodies of knowledge[3] that are stored in long term memory and are functionally used in most of the higher cognitive competences, where the relevant competences are such things as categorization, induction, deduction, concept combination, and planning (Machery 2009; Weiskopf 2009). While there are alternate notions of a concept,[4] in this paper we will understand concepts in the widely understood sense in cognitive science and philosophical psychology. Here, concepts are the constituents of thought and mental representations used in most of the higher competences.

Concepts are theoretical constructs in psychology that play a pivotal role in explaining higher acts of cognition. Moral concepts and their structures are in significant part responsible for how we perform competences in moral cognition such as moral

[2] I follow normal convention in the concepts literature and capitalize all concepts.

[3] Throughout this paper, 'knowledge' will stand for an information-carrying mental state rather than the traditional philosophical understanding of true justified belief. In this respect, I follow standard convention in the concepts literature.

[4] Fodor's informational atomism theory of concepts is meant to provide a theory of content for concepts and not necessarily provide a view of how concepts partake in the higher competences (Fodor 1998). Providing a theory of content and a theory of how concepts play a functional role in higher cognition can be seen as being in principle two different projects, although one may pursue both projects. Also, the notion of a concept used here differs from the idea of a platonic concept that is an abstract object rather than a mental representation. While platonic concepts are focused on the metaphysically correct features of a category, the interest in this paper will be on epistemic mental representation concepts in the minds of human beings that may change over time and may be incorrect. Finally, some psychological concepts can be personal states and others can be subpersonal states. Insofar as personal states have the function of being intentionally used and subpersonal states do not have this function, psychological concepts can be either of the two kinds of states.

categorization, decision-making, planning, analogical reasoning, and induction. However, very few concept theorists have worked on the nature of abstract moral concepts. No experiments have been run explicitly on the theory-theory for moral concepts, and it is not even clear what theory knowledge is in the moral concepts domain. Despite the central importance of moral concepts for moral cognition, surprisingly, only a few moral psychologists explicitly have worked on this project.[5] The causal judgment literature in moral psychology that examines what mental states influence moral decision-making rarely discusses the concepts literature, and the causal judgment literature generally does not explicitly examine the structure of moral concepts. The causal judgment literature in moral psychology generally does not discuss the main player in moral cognition and moral decision-making in the psychological domain; namely, moral concepts. However, as we shall later see, since concepts and their psychological structures are defined as influencing decision-making and the causal judgment literature examines what mental states influence moral decision-making, I will use the causal judgment literature in moral psychology to draw certain conclusions about the structure of some moral concepts.

Although when asked, concept theorists likely will not deny the possibility that moral concepts might have theory structure, it has yet to be stated and proven in the literature that moral concepts indeed do have theory structure. Several studies have demonstrated that some abstract concepts may have different structures than those commonly found in concrete concepts (Hampton 1981; Barsalou et al. 2005; Wiemer-Hastings et al. 2005). For instance, James Hampton, ran tests on eight different abstract concepts, such as BELIEF, SCIENCE, and CRIME, along with tests on some concrete concepts in order to determine whether the abstract concepts had prototype structure similar to the successful results of finding prototype structure in concrete concepts. Although we will elaborate on this theory later, the prototype view claims that concepts are constituted by prototypes or mental representations of the statistically frequent features of members of a class (Rosch and Mervis 1975). In Hampton's study, the results were a mixed bag where some abstract concepts did show prototype structure, but others did not. For example, SCIENCE and CRIME showed prototype structure while abstract concepts such as A BELIEF and AN INSTINCT, that may intuitively be thought to have prototype structure, as a matter of fact do not have such structure. Thus, the upshot from Hampton's and others' experiments is that we cannot safely presuppose that abstract concepts will have the same theoretical concept structure and cognitive processing as those for concrete

[5] Some exceptions are Jesse Prinz (2008), Stephen Stich (1993), Mark Johnson (1993), Paul Churchland (1989), and David Wong (2006).

concepts. As a matter of caution, we cannot draw conclusions about moral concepts solely based on the findings of concrete concepts. Therefore, further work is required in order to ascertain the structure of moral concepts, such as whether some of them have theory structure. I will put forth this further work in order to demonstrate the viability of the theory-theory for some moral concepts.

The Theory-Theory for Concrete Concepts

We will now elaborate on the theory-theory for concrete concepts in order to construct and get a proper sense of how the theory-theory may look like for moral concepts. The theory-theory of concepts is a view that emerged out of psychology in the 1980s, although in philosophy it has its roots in the likes of Locke (1689) and Quine (1977). The theory-theory of concepts states that concepts are themselves theories or mini-theories. Theories or mini-theories are certain mental representations. More specifically, they are scientific, hidden essence, causal law-like, functional, explanatory, and general or generic background knowledge about the extension of a concept and can explain such things as categorization in concrete concepts. For example, Edouard Machery, who is a proponent of this view *inter alia*, writes:

> Psychologists assume that laws, causal propositions, functional propositions…, and generic propositions…explain why things happen. Thus, a theoretical concept is supposed to store some nomological, causal, functional, and/or generic knowledge about the members of its extension. (2009, 101)

A description of these various kinds of theory mental representations will be discussed below in turn.

It will help to understand the theory view, especially for those who are not familiar with the concepts literature, by contrasting it with the prototype theory of concepts. As we shall later see, a major shortcoming of the prototype view is its inability to account for the additional knowledge that has been found to be stored in many concepts that accounts for some cases of classification. Recall that the prototype view claims that concepts are constituted by or just *are* prototypes or mental representations that refer to the statistically frequent features of members of a class. By the term 'constituted,' I am referring to the 'is' of identity. For example, Jesse Prinz, who is a proponent of this view *inter alia*, states that, "[M]any categories are associated with small sets of "typical" features. Typical features are ones that are diagnostic, statistically frequent, or salient. Unlike defining features, they are often contingent for category membership" (2002, 52).

Inspiration for the prototype view explicitly comes from Wittgenstein's notion of family resemblance, where members of a category may have one to several features or characteristics in common with each other, but zero or very few characteristics are common to all category members.[6] Prototype features are considered statistically frequent in that it is highly probable that a member of one's category will have them. Prototypes represent the superficial appearances of an object. Moreover, prototypes do not represent features that are necessary and sufficient conditions for determining membership. For example, one's prototype of DOG may be HAS HAIR, HAS FOUR LEGS, BARKS, PLAYS FETCH, and WAGS ITS TAIL. The features that are represented may be arrived upon based on all of one's previous experiences with particular dogs. Such prototypes can influence categorization decisions of whether something is or is not a dog. Furthermore, a dog may still be considered a dog even though it is hairless or even if it only has three legs.

On prototype views, features may be weighted more heavily than others. For example, *barks* and *plays fetch* may be weighed more heavily than *has hair*. When calculating how similar a potential or target member may be to a category, a prototype theory may take into account the number of features the instance may share with a category, or the instance's satisfaction of heavily weighted features, or both. Prototype theory is considered a similarity-based view because when an object or act is similar enough to the representation of summary or general features and passes a calculated similarity threshold, then the object falls within the class. Returning to the example, since my pet animal satisfies the importantly heavily weighted features of *barks* and *plays fetch*, along with several of the other features, it passes the similarity threshold and is categorized as a dog. When a token passes the similarity threshold of two or more categories, it is generally categorized in the class towards which it has the highest similarity score. At the same time, passing the similarity threshold for multiple categories can also beneficially explain the phenomena of ambiguous cases of membership, where some individuals may categorize an item as a member of two different classes.

The theory-theory attempts to address this issue of the superficial nature of the features represented by prototypes by claiming that background knowledge of the world rather than prototypes that are about superficial properties can play a role in higher acts of cognition. In other words, theories or mental representations of things like hidden essences, causal laws, and functions can determine how we make categorization

[6] There are various versions of the prototype theory and disagreements between different proponents of such views. However, I detail here the prototype theory as it generally may be understood.

decisions. Use of theory knowledge as well as other conceptual knowledge, such as prototypes, in cognition may be effortful, automatic, conscious, or subconscious.

In an experiment for the theory-theory, Frank Keil ran a study where participants were asked whether the animal in a given scenario is a horse or a cow (1989, 162). In the situation, there is an animal that is called a 'horse,' makes horse sounds, looks like a horse, is strapped with a saddle so people can ride on it, and eats oats and hay. The animal has all the superficial prototypical features of a horse. However, scientists run blood tests and x-rays on it, and they discover that its insides are actually the insides of a cow. In this experiment, Keil found that older children and adults perceived the scientists' discoveries as relevant for determining natural kind membership. These subjects relied not on superficial similarities but on folk biological theories of hidden essences to decide that the animal was really a cow despite its superficial horse appearances.

As an example of the importance and use of folk causal law knowledge in cognition, *being curved* is an equally typical prototype feature in bananas and boomerangs. However, subjects give more weight to this attribute in boomerangs rather than bananas because it is falsely believed that curvature is causally related to the boomerang's property of *if thrown, it will return back to the thrower* (Medin & Shobin 1988). Due to this relationship between the two features, it is thought that *being curved* is more required for a boomerang rather than a banana. Theories may provide other causal explanatory relations between superficial features of an object. As an example that is an oversimplification for the point of illustration, my FISH concept may be constituted by the prototypes: HAS FINS, HAS A TAIL, and SWIMS (Murphy & Medin 1985). Theoretical knowledge of fish provides the explanation of the relation between fish attributes since in order to properly swim, a fish needs fins and a tail. Furthermore, participants believe that the hidden essence of a natural kind is generally causally responsible for the superficial features of the kind. For instance, many believe that the hidden essence of human beings is responsible for why we have the typical observable properties that we do. Here, most theory-theorists usually do not necessarily deny that one may have in mind superficial features when representing a class (Medin & Shobin 1988; Murphy & Medin 1985; Carey 2009), but they do emphasize the importance of such things as folk causal law knowledge or theories in providing the underlying explanation to such features as well as in deciding what weight such features may possess.

Theory-theorists also hold that there are domain differences for types of knowledge where different ontological domains contain different types of central beliefs. For example, while natural kinds are believed to have hidden essences, the analogue for artifact kinds generally is intended function. For example, Lin and Murphy ran an experiment where they first described and showed pictures of certain artifacts from foreign countries (1997). One such item is a *tuk*. A tuk is a hunting tool that is a stick with a special handle on one end that protects the wielder's hand from animal bites. On the other end of the stick is a noose that goes around the head of the animal. The function of the tuk is to be able to control an animal by placing the noose over its head. After informing participants about what a tuk is and the function it performs, the experimenters showed participants a picture of what looks like a tuk minus only the special handle. When asked to categorize the item, participants categorized it as a tuk. When shown a picture of what looks like a tuk minus only the noose, subjects did not categorize it as a tuk. This suggests that functional knowledge plays a role in categorization, where participants did not categorize the latter item as a tuk because it could not perform the tuk's function.

Moreover, theory knowledge may not only contain knowledge of hidden essences, theoretical entities, and causal laws, but they may also contain general background knowledge. Theory knowledge need not be restricted to those kinds of knowledge that are of properties that are related to the structure of scientific theories, properties such as causal laws and essences. For example, Murphy and Medin claim that most people think the feature of *flammable* is a quality of wood rather than paper money even though both wood and paper money are flammable (1985). The reason behind this is that we have general background knowledge about the world concerning human activity where wood is used for burning fires and paper money is mostly used for economic purposes in which its flammability plays no role. On their view, this knowledge still counts as a theory that influences feature attribution for classes even though it may not be about a causal law or hidden essence.[7] For, such general background knowledge about human activity in principle still does provide an *explanation* of why we may attribute certain features to certain classes.

The Theory-Theory for Moral Concepts

In this section, we will strike new ground in detailing how the theory view will look for moral concepts. In order to properly discuss how we may make the appropriate changes to the theory-theory in order to account for the moral domain and moral concepts, it will help to discuss the prototype theory of moral concepts. Just as we used

[7] In this respect, their understanding of the theory-theory differs from the likes of Gopnik and Meltzoff who draw a much tighter connection between theory conceptual structure and properties that are related to the structure of scientific theories.

the prototype theory above to help illustrate how the theory view posits background knowledge conceptual structures to concrete concepts that underlie superficial prototypes, we will likewise use the prototype theory of moral concepts in order to help illustrate the import and nature of the theory view for moral concepts.

If prototype theory is viable for moral concepts, then the prototype for the moral concept RIGHT ACTION for some individual may be BEING GENEROUS TO OTHERS, HELPING THE HOMELESS IS THE RIGHT THING TO DO BECAUSE IT BENEFITS THOSE IN NEED, PREVENT HARM, DOES NOT BREAK LAWS, and EXHIBITS FRIENDLINESS (Walker & Pitts 1998; Walker & Hennig 2004; Park 2013). Such representations, when understood as features of members of a class, are not necessary and sufficient conditions. As we can see, prototypes may be about such things as general features of moral situations, virtues, reasons for action, and basic moral principles or rules. For example, when a person points out to another an instance where a stranger is helping homeless people that such is a case of moral rightness, the mentally represented reason or justification for action that HELPING THE HOMELESS SO LONG AS ONE IS NOT IN POVERTY ONESELF IS THE CORRECT THING TO DO BECAUSE IT BENEFITS THOSE IN NEED may now be a candidate to be a constituent component of this listener's prototype of RIGHT ACTION based on further particular experiences. The abstraction of such features may be based on personal experiences and moral education. Now, for the individual in question, the virtue friendliness may carry less weight for this individual as compared to heavily weighted features such as the principles *prevent harm* and *does not break laws*.

Concerning the theory-theory as applied to ethical concepts, ethical concepts may themselves be theories that have as components such things as knowledge about master moral principles from which other moral principles generally may be thought to be inferred and explained. For example, the concept RIGHT ACTION may be constituted by normative theoretical information akin to divine command theory. The theory ACT IN ACCORDANCE WITH THE PRINCIPLES MANDATED BY GOD can be a component of RIGHT ACTION.

As previously stated, prototypes are about features that may be such things as moral principles, reasons for action, and virtues. However, the theory-theory components are more about master moral principles from which other moral principles, reasons, and virtues generally may be thought to be inferred and explained. For example, from divine command theory, where one must obey those laws mandated by God, one may arrive upon principles such as *do not lie* and *do not steal*. One may adhere to these moral principles based on one's background belief in the ethical theory of divine command theory. Moreover, for this person, divine command theory explains why we must not lie and steal. On the other hand, one may have a virtue ethics master moral principle in mind such as EXEMPLIFY THOSE VIRTUES THAT THE VIRTUOUS PERSON HAS from which one may infer the proper virtues, such as KINDNESS, HONESTY, GENEROSITY, and PATIENCE. Moreover, this theory knowledge or master principle explains why the particular individual adheres to the group of virtues that she does. Also, a person may have in mind the universalizability principle as theory knowledge to infer maxims. The mentally represented deontological master principle OBEY THAT MAXIM THAT ONE CAN WILL TO BE A UNIVERSAL LAW can be used by a particular person to infer and explain the set of maxims this individual holds, such as DO NOT LIE and DO NOT STEAL.

Also, the act utilitarian Greatest Happiness Principle, where one must perform that action that leads to the greatest happiness for the greatest number, initially appears to be only about generating verdicts on acts rather than deriving other principles. However, act utilitarianism may not produce formal pithy rules or principles but it still relies on reasons for action to arrive upon its act-based conclusions. As a utilitarian, in self-defense one may believe that one ought to take the life of a murderous assailant. However, this may be based on the specified reason that in regards to overall happiness, killing others is wrong, but in acts of self-defense, killing others is justifiable. Such a reason may be thought to stem from and be explained by utilitarianism in that it is under the eye of the Greatest Happiness Principle that such a reason or justification is formed and used to determine what act one should follow. Thus, act utilitarianism may also be thought to be a theory or master principle from which reasons or considerations that count in favor of something are based.

While the theory-theory is a distinct view from the prototype theory based on the specific kind of knowledge it posits concepts as containing, as we can see, there is an intimate link between the theory-theory and the prototype theory for moral concepts. Ethical theory knowledge provides an underlying explanation for why one holds the moral prototypes that one does. This is just like the intimate link between both concept theories in the concrete concept realm, where as shown in our previous example, general explanatory background knowledge is the reason why the folk attribute the prototypical property of *flammable* to wood but not paper money. Also, recall that Murphy and Medin showed how theory knowledge need not be restricted to those kinds of knowledge that are of properties related to the structure of scientific theories. Hence, theory knowledge for moral concepts also need not be about properties related to the structure of scientific theories.

It is knowledge about these master principles that belong to the theory-theory of moral concepts, while the moral principles, reasons, and virtues that may be thought to be inferable from master principles belong more in the domain of the prototype view. The reason why this is the case is that such master principles are those that ostensibly underlie the inferential principles, virtues, and reasons for action at a deeper level of theoretical abstraction just as, for example, folk biological theories of hidden essences underlie the superficial features of a biological natural kind. Second, and a related point, is that just as biological theories may be about *explanatory relations* between superficial features, master moral principle knowledge may provide an explanatory link between the inferential principles, virtues, or reasons. In this respect, inferential principles, reasons, and virtues may be thought of as being superficial. Meanwhile, master principles are more theoretical and lie at the deepest explanatory level. For example, *do not kill* and other rules such as *do not lie* may be ultimately explained and unified by divine command theory for a particular person. Represented master moral principles are theory knowledge precisely because they provide an underlying explanation for why one holds the inferential principles, virtues, or reasons for action that one does. This is just like how some theory knowledge of a class in the concrete concepts domain provides an underlying explanation of one's prototypes of that class.

What I have provided thus far for an explanation of the structure of the theory-theory for moral concepts may not exactly parallel the structural components of the theory-theory for concrete concepts given the differences between the abstract moral and concrete concept domains. However, in allying some theories with knowledge of master moral principles, we can see that several important similarities as just discussed exist between the given moral and concrete theory structures to warrant drawing the distinction between moral concept prototype and theory structural components in this manner.

Also, causal moral law knowledge about an agent's intentions for action may be theory components of moral concepts in that they may take part in the higher cognitive competences related to ethical matters. Just as natural kind concepts may contain causal law knowledge, moral concepts may also be constituted by causal moral law knowledge. For instance, in moral cognition, one may have a representation of a causal moral law or principle such as IF AN AGENT'S CAUSAL MOTIVATIONS ARE WRONG, THEN THE AGENT'S ACTION USUALLY IS WRONG. This may constitute one's WRONG concept. It may also underlie and be responsible for the prototype constituent of one's WRONG concept, LACKS MORAL WORTH, to carry significant weight and importance. If one does not perform an action with the proper motivations and intentions, then the act lacks moral worth, and this may lead to the greater chance that the act is classified as a wrong act. Given the above causal moral law knowledge, LACKS MORAL WORTH may be held to contain significant weight and importance in one's WRONG concept. This is somewhat analogous to the theory-theory for concrete concepts in which one's BOOMERANG concept may be constituted by theory knowledge of the causal law-like principle IF THROWN, IT WILL RETURN BACK TO THE THROWER. Recall that this theory knowledge likewise underlies and impacts the weight carried by the prototype constituent of one's BOOMERANG concept, BEING CURVED. BEING CURVED is a more heavily weighted prototype component for BOOMERANG rather than BANANA because the causal principle knowledge contained in participants' BOOMERANG concept confers such additional weight.

In summary of this section, I have developed the theory-theory for moral concepts. Theory knowledge for moral concepts includes such things as ethical theory knowledge of master principles and causal moral law knowledge. Just as the use of hidden essence knowledge in the categorization of natural kinds provides evidence for the theory view in respect to such categorization, the use and influence of ethical theory and causal moral law knowledge in moral cognition when making moral categorization decisions of what is morally right and wrong will provide evidence for the theory view in respect to such tasks. Conceptual structures that constitute a concept, such as prototypes and theories, are posited as playing a functional or causal role in higher acts of cognition, such as in categorization and decision-making. In the subsequent section, we will first examine studies showing the use of ethical theory knowledge at times in moral categorization. Next, we will discuss several experiments that demonstrate the use of causal moral law knowledge at times in moral categorization. These various sets of studies to be examined each independently demonstrate the viability of the theory-theory for moral concepts.

As is becoming widely accepted in the concrete concepts literature (Prinz 2002; Murphy 2004; Machery 2009; Weiskopf 2009), I perfectly allow for the possibility that other kinds of knowledge other than theory knowledge can be used at times in cognition. For example, it may be the case that emotions are so used, which will lead to a view in which moral concepts are at least in part constituted by emotions. Prototype and other kinds of knowledge may also in part constitute our moral concepts, where different kinds of concept constituents can be used together or separately in moral cognition in different contexts. For instance, in one case of decision-making I may rely only on theory knowledge, and in another situation I may use only prototypes and emotions.

We will now examine several studies demonstrating that theory knowledge is also at times used in moral categorization. Although there may be other kinds of knowledge used in moral cognition, due to space concerns, the focus here is precisely on the explication of the theory view in the moral concepts domain and providing evidence that theory knowledge is actually used at times in moral categorization.

Evidence for the Theory-Theory of Moral Concepts

Now that we have clarified what kind of knowledge is theory knowledge for moral concepts, we will examine well-known studies that are not explicitly designed by their authors to draw moral concept constitution conclusions, but they can indeed be used as evidence for the use of theory knowledge in moral categorization of what is morally right and wrong. As previously stated, there are no explicit studies on the theory-theory for moral concepts in the concepts literature or in moral psychology. However, there are numerous experiments in moral psychology that provide evidence for the theory-theory of moral concepts even though the relevant literature does not at all discuss how such studies can be used to draw structural conclusions on moral concepts. The relevant connection has not been made in the relevant studies. In other words, there are several studies showing that at times, theory knowledge is used in moral categorization. However, in the literature, the connection is not drawn that since concepts are functionally defined, such studies provide evidence for the theory-theory of moral concepts. Another reason why this connection has not been made is that no one to this point has stated in the literature what theory knowledge even is for moral concepts. As we shall see, the use of theory knowledge in moral cognition is just like how in Keil's previously discussed study, participants used theory knowledge of hidden essences in order to make their categorization judgment that the given animal is really a cow instead of a horse.

Joshua Greene, et al. ran a cognitive load study where subjects were filling out moral questionnaires on a computer (2008). They were presented with "high conflict" moral dilemmas in which subjects were asked whether it is appropriate to harm another individual in order to save several lives. One example of a high conflict dilemma that was used is the crying baby case:

> Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Outside you hear the voices of soldiers who have come to search the house for valuables.

> Your baby begins to cry loudly. You cover his mouth to block the sound. If you remove your hand from his mouth his crying will summon the attention of the soldiers who will kill you, your child, and the others hiding out in the cellar. To save yourself and the others you must smother your child to death.
>
> Is it appropriate for you to smother your child in order to save yourself and the other townspeople? (Greene et al. 2008, 1147–1148)

While answering such questions on moral dilemmas, numbers continually stream across the bottom of the screen and participants have to press a button when they see the number five. The result is that subjects selectively had a longer reaction time when making utilitarian judgments under the cognitive load as opposed to having no cognitive load, but there was no increase in reaction time for non-utilitarian judgments under cognitive load. This study provides causal rather than correlational support that utilitarian theory knowledge influences some moral judgments for some people because the longer reaction time suggests that the cognitive load of having to press a button when seeing the number five interferes with a controlled cognitive process, such as some kind of cost-benefit analysis reasoning process, whereas the cognitive load should have no effect on a fast automatic process.[8] This conclusion is further buttressed by neuroimaging

[8] Some of the provided scenarios to participants use the famed trolley problems. While the likes of Greene and Haidt interpret the lever and footbridge cases as that between utilitarianism and deontology (Haidt 2012; Greene 2013), others such as Mikhail interpret both cases as involving the use of the Doctrine of Double Effect (Mikhail 2007). However, as is well documented, Greene and Haidt have pointed out that the Doctrine of Double Effect interpretation is false given results from a loop variant scenario in which participants will use a person as a means to save many lives (Haidt 2012; Greene 2013). In other words, subjects do not consistently abide by the Doctrine, and the opposite predictions from the Doctrine interpretation are borne out. Hence, I interpret studies reliant on the trolley problems to pit utilitarianism knowledge versus deontological knowledge. To note, even if the Doctrine was at work, its use in cognition would still provide evidence for the theory-theory; a concept constitution conclusion that is not foreseen nor anticipated in the writings of Mikhail. Mikhail, like other moral psychologists, fails to draw the connection in his works.

Guy Kahane has argued in the trolley problems that the supposed evidence for the use of utilitarianism knowledge is really support for the use of deontological theory knowledge (2012). Yet, one major problem with Kahane's thesis is that 44 independently-run cross-cultural studies have demonstrated that there is extremely little correlation between moral judgments and deontological thinking, but there is a robust correlation between moral judgments and utilitarian thinking (for a summary of the studies, see Snarey 1985). While this robust correlation provides even further support for the particular theory-theory conclusion that utilitarian knowledge is used in moral cognition, without a general correlation between studies that show a strong correlation between the making of utilitarian judgments and activation in the dorsolateral prefrontal cortex (Greene et al. 2004). This brain region is known for such things as complex planning, deductive/inductive reasoning, and long-term economic decision-making. The above provides evidence for the viability of the theory-theory for moral concepts.

Mendez and company have shown that frontotemporal dementia patients who have intact reasoning capacities but who have blunted affect or severely diminished emotions, tend to make the same utilitarian moral judgments as compared to normal subjects on the same moral scenarios (2005). Moreover, Koenigs, Young, and company (2007) as well as Ciaramelli and colleagues (2007) have demonstrated that patients with lesions to the VMPFC who have blunted affect but intact reasoning capacities also tend to make the same utilitarian judgments as normal participants on certain moral vignettes. Furthermore, the patients and the normal subjects displayed activation in the dorsolateral prefrontal cortex, like in Greene et al.'s above study. Since the patients have blunted affect and make the same utilitarian judgments as normal persons, this provides evidence that cognitive utilitarian reasoning is being used by normal agents in order to make normal moral judgments in certain cases. These studies show that what at least in part influences moral categorization in such cases is cognitive knowledge, not emotions. This is one important step for the establishment of the theory-theory since the theory view is a cognitivist one. Moreover, this cognitive knowledge involves a utilitarian calculation. In other words, it is knowledge of a master moral principle and is therefore, not prototype knowledge. Thus, this provides evidence for the theory-theory for moral concepts in that theory knowledge is being used at times in ordinary moral categorizations by normal participants just as theory knowledge of the function of a tuk was used by participants to judge that an object is not a tuk since the given object cannot perform the function of a tuk.

We now will turn to experiments that demonstrate that causal moral law knowledge at times influences moral categorization. This will provide additional independent evidence for the viability of the theory-theory for moral concepts. There is good evidence

moral judgment and deontological thinking, there is no causation; there is no general causal influence of deontological theory in moral categorization for normal subjects. However, even if Kahane is correct, then the trolley problems provide evidence for the use of deontological *theory* knowledge rather than utilitarian-like theory knowledge. Therefore, we still get our general desired conclusion, and the trolley experiments still provide evidence for the use of some kind of ethical theory knowledge in moral categorization. To note, Kahane likewise fails to draw the connection between moral psychology and the concepts literature in his writings.

that causal moral law knowledge is involved at times in moral categorizations (Cushman 2008; Young & Saxe 2008). As a moral categorization example, Paharia, et al. gave a group of subjects the following situation:

> A well-known real estate developer, X, owned a piece of property they wished to construct new housing units on. The property contained some health-threatening toxic substances that would require a substantial amount of clean-up, and was worth $50 million dollars as is. It would require $30 million to fully clean the land, but the value would only go up to $60 million. [The housing developer decided to only invest $12 million in a 40% clean up effort, and the value of the land went up to $54 million. They built housing units on the land, all of which have now been sold.] (Paharia 2009, 136)

Meanwhile, another group of participants received the same vignette except the text in the bracket was replaced with: "The housing developer sold the land to a lesser-known developer, Y, without cleanup. The lesser-known developer invested no money in any clean up effort and built housing units on the land, all of which have now been sold (136)." The experimenters discovered that participants judged developer X to be less unethical in the second situation when they sold the land to developer Y as compared to the first situation in which X directly rather than indirectly caused the property to not be fully cleaned up. Notice that they judged X to be less culpable in the situation where there was no clean up rather than in the case where there was a 40% clean up. Furthermore, in a third scenario, the experimenters found that these results held even when in later studies it was explicitly stated to subjects that agent Y was an instrument of agent X, contracted to do its bidding. These studies provide support for the theory-theory in that the best explanation for why these series of judgments are made in the three different scenarios is that participants take into account whether an agent is a direct or indirect cause for action (or inaction) when making moral categorizations. If only emotions such as anger or, for that matter, any kind of knowledge influences judgments without any conjoint influence whatsoever from the above causal knowledge regarding whether the agent is a direct or indirect cause of an action, then we should expect participants to claim that in all three situations, agent X is equally culpable or that perhaps X is more culpable in the second and third scenarios since there is no clean up whatsoever. However, the fact that subjects generally judge X to be most culpable in the first scenario, which is the only circumstance where X is directly responsible, strongly suggests that some kind of a general causal moral principle is in play and is being used in categorization in order to properly explain and make sense of the discrepancy in judgments for the above three scenarios. This causal moral law knowledge places more culpability on those who directly causally influence an immoral outcome as contrasted with being an indirect cause of an immoral outcome. This causal principle knowledge is at work for many subjects when making certain judgments, regardless of whether emotions or some other kind of conceptual constituents are jointly also in play working together with the causal principle in moral cognition. That this knowledge is not an emotion is one important step in establishing the viability of the theory-theory for moral concepts since the theory view is a cognitive account of concepts. Furthermore, that this knowledge is of a causal moral law establishes that it is theory rather than prototype knowledge. Therefore, this study provides evidence for the use of theory knowledge in moral categorization.

Finally, Young, et al. discovered that causal moral law knowledge of an agent's causal intentions to act influences moral categorization (Young et al. 2010). In other words, an agent's causal intentions which influence the agent's behavior affect one's moral judgment of the agent. In this study, the experimenters used trancranial magnetic stimulation (TMS) in order to disrupt the neural activity in the right temporoparietal junction (RTPJ) before and during moral judgment. There has previously been shown to be a correlation between activation in the RTPJ and when a participant reads about an agent's causal intentions for action in certain moral contexts (Young et al. 2007). What they found was that in attempted harm cases, where the agent causally intends to do harm but fails to bring about the negative consequences, TMS to the RTPJ causes subjects to judge the agent's attempted harm as being less morally forbidden and more morally permissible as compared to participants who received TMS to a control site. Due to the nature of attempted harm cases, this suggests that TMS to the RTPJ affects the ability of subjects to fully account for the agent's causal intentions in making moral judgments and that assessing an agent's causal intentions does play a role in influencing typical normal moral judgments. This study demonstrates that in many normal moral judgments for normal subjects who are not receiving TMS to the RTPJ, such participants take into account causal knowledge about an agent's intentions and what causally motivates the agent to perform a given action. Normal participants have causal moral law knowledge that if an agent's causal motivations for action are bad, then the agent's action is still wrong even though the agent fails to bring about the bad consequences. This causal moral principle knowledge impacts their moral decision-making. Given that the causal knowledge of an agent's intentions for action affects the gravity of moral judgments, this provides evidence for the viability of the theory-theory for moral concepts in light of categorization. In this section, I have discussed several different sets of studies, where each set can independently show that some people at least in part have theory knowledge stored in their moral concepts. My conclusion that some moral concepts have theory structure is an advancement beyond the current literature.

Conclusion

I purport to have provided two main contributions to the concepts literature in philosophy of cognitive science. I have first shown how the theory-theory will look like for moral concepts and have shown what kinds of knowledge are theory knowledge in the moral domain. Currently, the concepts literature does not state what counts as theory knowledge for moral concepts. Second, while I perfectly leave open the possibility that our moral concepts may also store different kinds of knowledge, such as prototypes, that are used individually or conjointly with other kinds of knowledge or information-carrying mental states in moral cognition depending upon the circumstances, I freshly have demonstrated that for moral categorization of what is morally right and wrong, there is in certain circumstances evidence for the theory-theory of moral concepts. Currently in the concepts field, it has not been stated nor explicitly shown that the theory view applies to the moral concepts domain for moral categorization. The relevant connection between the concepts and moral psychology fields has not been made in the literatures. While I perfectly leave open the possibility that theory knowledge may also apply to other different domains of concepts, my focus here is exclusively on moral concepts.

My conclusions are a non-trivial matter since as mentioned above, numerous experimental studies demonstrate that one cannot draw structural concept conclusions on abstract concepts based solely on data from concrete concepts. One cannot infer that moral concepts in part have theory structure based on the evidence for the theory-theory in the concrete concepts domain. Moreover, it is also non-trivial in that it has never been stated in the literature what theory knowledge even is for moral concepts; without which no one lucidly can claim that the theory view is viable for moral concepts. Furthermore, as we have discussed, several moral psychologists independently have had to devise and run numerous clever studies to sufficiently prove that the relevant knowledge is in fact used in moral cognition. The establishment of the theory view for moral concepts is hardly trivial.

A third contribution to the concepts field is that the above work on the theory view for moral concepts may open up new lines of future empirical work on explicitly examining what other kinds of theory knowledge may be stored in our moral concepts. For example, there are a variety of different ethical views in philosophy, such as

Aristotelian virtue ethics and Kantian deontological theory. Do some people store virtue ethical theory knowledge in their moral concepts or in part other theory knowledge? Also, while I have merely claimed here that *some* moral concepts have theory structure, there are many different types of moral concepts. More studies can be run on moral concepts not examined here, such as on HONESTY, INTEGRITY, and GREED, to see whether they also may have theory structure.

Although there presently is insufficient evidence to examine whether theory knowledge for moral concepts plays a role in other higher acts of cognition, such as in concept combination, induction, and analogical reasoning,[9] I have shown how experiments that were not explicitly designed to test for the theory-theory of moral concepts actually can be used to demonstrate that theory moral knowledge is used at times in moral categorization just as theory knowledge of hidden essences is used at times by many to classify an object as being a cow rather than being a horse. I leave further examination of the full extent of the use of theory knowledge in moral cognition for a later time. Recall that such theory knowledge can be used individually or conjointly with other kinds of knowledge in moral cognition depending on the context. There even may be situations where theory knowledge is not being used in particular cases of moral decision-making.

References

Barsalou, L., and K. Wiemer-Hastings. 2005. "Situating Abstract Concepts." In *Grounding Cognition*, edited by D. Pecher and R. Zwaan, 129–163. Cambridge: Cambridge University Press.

Carey, Susan. 1985. *Conceptual Change In Childhood*. Cambridge, MA: The MIT Press.

Carey, Susan. 2009. *The Origin of Concepts*. Oxford: Oxford University Press.

Ciaramelli, E., M. Muccioli, E. Ladavas, and G. di Pellgrino. 2007. "Selective Deficit in Personal Moral Judgment Following Damage to Ventromedial Prefrontal Cortex." *Social Cognitive and Affective Neuroscience* 2: 84–92.

Churchland, Paul. 1989. *A Neurocomputational Perspective*. Cambridge, MA: The MIT Press.

[9] If moral concepts do use theory knowledge for categorization, we should expect that such knowledge should be able to partake in the other higher competences just as well as the theory knowledge in concrete concepts can.

Cushman, Fiery. 2008. "Crime and Punishment: Distinguishing the Roles of Causal and Intentional Analyses in Moral Judgment." *Cognition* 108: 353–380.

Fodor, J. 1998. *Concepts*. Oxford: Oxford University Press.

Frei, Jennifer, and Phillip Shaver. 2002. "Respect in Close Relationships: Prototype Definition, Self-Report Assessment, and Initial Correlates." *Personal Relationships* 9: 121–39.

Gopnik, Alison and Andrew N Meltzoff. 1997. *Words, Thoughts, and Theories*. Cambridge, MA: The MIT Press.

Greene, J, L.E. Nystrom, A.D. Engell, J.M. Darley, and J.D. Cohen. 2004. "The Neural Bases of Cognitive Conflict and Control in Moral Judgment." *Neuron* 44: 387–400.

Greene, Joshua, Sylvia Morelli, Kelly Lowenberg, Leigh Nystrom, and Jonathan Cohen. 2008. "Cognitive Load Selectively Interferes with Utilitarian Moral Judgment." *Cognition* 107*:* 1144–54.

Greene, Joshua. 2013. *Moral Tribes.* New York: The Penguin Press.

Haidt, Jonathan. 2012. *The Righteous Mind*. New York: Vintage Books.

Hampton, James. 1981. "An Investigation of the Nature of Abstract Concepts." *Memory & Cognition* 9 (2): 149–156.

Johnson, Mark. 1993. *Moral Imagination*. Chicago: The University of Chicago Press.

Kahane, Guy. 2012. "On the Wrong Track: Process and Content in Moral Psychology." *Mind & Language* 27: 519–545.

Keil, Frank C. 1989. *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: The MIT Press.

Koenigs, M., L. Young, R. Adolphs, D. Tranel, F. Cushman, M. Hauser, A. Damasio. 2007. "Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgments." *Nature* 446: 908–911.

Lin, E. and G. Murphy. 1997. "The Effects of Background Knowledge on Object Categorization and Part Detection." *Journal of Experimental Psychology* 50A: 25–48.

Locke, John. (1689) 1996. *An Essay Concerning Human Understanding*. Edited by Kenneth P. Winkler. Indianapolis, IN: Hackett Publishing Company, Inc.

Machery, Edouard. 2009. *Doing Without Concepts*. Oxford: Oxford University Press.

Medin, D. and E. Shoben. 1988. "Context and Structure in Conceptual Combination." *Cognitive Psychology* 20: 158–90.

Mendez, M.F., E. Anderson, and J.S. Shapria. 2005. "An Investigation of Moral Judgment in Frontotemporal Dementia." *Cognitive and Behavioral Neurology* 18 (4): 193–7.

Mikhail, John. 2007. "Universal Moral Grammer: Theory, Evidence and the Future." *Trends in Cognitive Sciences* 11 (4): 143–152.

Murphy, Gregory. 2004. *The Big Book of Concepts*. Cambridge, MA: The MIT Press.

Murphy, G. and D. Medin. 1985. "The Role of Theories in Conceptual Coherence." *Psychological Review* 92: 289–316.

Paharia, N., K. Kassam, J. Greene, and M. Bazerman. 2009. "Dirty Work, Clean Hands: The Moral Psychology of Indirect Agency." *Organizational Behavior and Human Decision Processes* 109 (2): 134–141.

Park, John J. 2013. "Prototypes, Exemplars, and Theoretical & Applied Ethics." *Neuroethics* 6: 237–247.

Prinz, Jesse. 2002. *Furnishing the Mind*. Cambridge, MA: The MIT Press.

Quine, W.V.O. 1977. "Natural Kinds." In *Naming, Necessity, and Natural Kinds*, edited by S.P. Schwarz, 155-175. Ithaca, NY: Cornell University Press.

Rosch, Eleanor and Caroline Mervis. 1975. "Family Resemblances: Studies in the Internal Structure of Categories." *Cognitive Psychology* 7: 573–605.

Smith, Kyle, Seyda Smith, and John Christopher. 2007. "What Defines the Good Person?" *Journal of Cross-Cultural Psychology* 38: 333–360.

Snarey, John. 1985. "Cross-Cultural Universality of Social-Moral Development: A Critical Review of Kohlbergian Research." *Psychological Bulletin* 97: 202–32.

Stich, Stephen. 1993. "Moral Philosophy and Mental Representation." In *The Origin of Values*, edited by M. Hechter, L. Nadel, and R. Michod, 215–228. New York: Aldine de Gruyer.

Walker, Lawrence and Russell Pitts. 1998. "Naturalistic Conceptions of Moral Maturity." *Developmental Psychology* 34: 403–418.

Walker, Lawrence and Karl Hennig. 2004. "Differing Conceptions of Moral Exemplarity: Just, Brave, and Caring." *Journal of Personality and Social Psychology* 86: 629–647.

Wiemer-Hastings, K., and X. Xu. 2005. "Content Differences for Abstract and Concrete Concepts." *Cognitive Science* 29: 719–736.

Weiskopf, Daniel. 2009. "The Plurality of Concepts." *Synthese* 169: 145–173.

Wong, David B. 2006. *Natural Moralities*. Oxford: Oxford University Press.

Young, L., F. Cushman, M. Hauser, and R. Saxe. 2007. "The Neural Basis of the Interaction between Theory of Mind and Moral Judgment." *Proceedings of the National Academy of Sciences of the United States of America* 104: 8235–8240.

Young, L., and R. Saxe. 2008. "The Neural Basis of Belief Encoding and Integration in Moral Judgment." *NeuroImage* 40 (4): 1912-1920.

Young, L., J. Camprodon, M. Hauser, A. Pascual-Leone, and R. Saxe. 2010. "Disruption of the Right Temporoparietal Junction with Transcranial Magnetic Stimulation Reduces the Role of Beliefs in Moral Judgments." *PNAS* 107 (15): 6753–6758.