Cognition 113 (2009) 80–92

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

The morality of harm

Paulo Sousa *, Colin Holbrook, Jared Piazza

Institute of Cognition and Culture, Queen's University, Belfast, 2-4 Fitzwilliam Street, BT71NN Belfast, UK

ARTICLE INFO

Article history:
Received 20 February 2008
Revised 8 December 2008
Accepted 19 June 2009

Keywords:
Morality
Harm
Transgression
Deontic reasoning
Social cognition

ABSTRACT

In this article, we discuss the range of concerns people weigh when evaluating the acceptability of harmful actions and propose a new perspective on the relationship between harm and morality. With this aim, we examine Kelly, Stich, Haley, Eng and Fessler's [Kelly, D., Stich, S., Haley, K., Eng, S., & Fessler, D. (2007). Harm, affect, and the moral/conventional distinction. *Mind and Language, 22*, 117–131] recent claim that, contrary to Turiel and associates, people do not judge harm to be authority independent and general in scope in the context of complex harmful scenarios (e.g., prisoner interrogation, military training). In a modified replication of their study, we examined participants' judgments of harmful actions in these contexts by taking into account their explanations for their judgments. We claim that both in terms of participants' judgments and rationales, the results largely confirm our hypothesis that actions involving harm *and injustice or rights violation* are judged to be authority independent and general in scope, which is a modification of Turiel's traditional hypothesis.

1. Introduction

Our main objective in this article is to discuss how people conceptualize the relation between moral wrongdoing and harm (in the sense of pain or, more broadly, suffering). We address questions such as: what are the components of the concept *moral transgression involving a person being subjected to harm*? Are harmful actions always understood as transgressions? Are they always understood as moral transgressions? To address these issues, we examine recent criticisms raised by Kelly, Stich, Haley, Eng, and Fessler (2007) against Turiel and associates (e.g., Nucci, 2001; Smetana, 1993; Tisak, 1995; Turiel, 1983, 2002), analyze qualitative evidence coming from an extended replication of Kelly et al.'s study, and elaborate an argument sketched elsewhere on the conceptual relation between moral wrongdoing and harm (see Sousa, 2009).

Over the years, Turiel and associates have accumulated a large body of evidence suggesting that people differentiate moral transgressions from conventional transgressions along several conceptual dimensions, the most important for our discussion being *authority contingency* and *generality*: moral transgressions are authority independent (i.e., their wrongness is not cancelable by the decision of any authority) and general in scope (i.e., their wrongness extends to different places and times), while conventional transgressions are authority dependent and local in scope. In other words, in the Turiel tradition, authority independence and generality comprise the major criteria that specify a transgression as a *moral* transgression. We shall call these criteria, following Kelly et al. the "moral signature." Our focus is on the hypothesis that transgressions involving a person being subjected to harm evoke the moral signature.

The moral/conventional task is the basic methodology utilized by the Turiel tradition. Harmful transgressions such as *a child pulling another child's hair* and conventional transgressions such as *a boy wearing nail polish* are presented in a random order to each participant in the task. They are not described as transgressions (moral or conventional), but simply as something someone does. For each action, a sequence of questions is then posed. We include

* Corresponding author.
E-mail addresses: psousa@qub.ac.uk (P. Sousa), cholbrook01@qub.ac.uk (C. Holbrook), jpiazza02@qub.ac.uk (J. Piazza).

0010-0277/$ - see front matter © 2009 Published by Elsevier B.V.
doi:10.1016/j.cognition.2009.06.015

here only those questions that are relevant to our discussion (*cf.* Nucci & Turiel, 1978; Smetana, 1981, 1985, 1986; Smetana & Braeges, 1990; Smetana, Schlagman, & Adams, 1993; Weston & Turiel, 1980). One standard version of the moral/conventional task proceeds as follows:

(Action Scenario)
1. 'Permissibility' probe
Is it OK for X to do A? YES NO
2. Justification probe
Explain your answer...
3. Authority contingency probe
Now, what if an authority says that it is OK to do A. Would it be OK for X to do A? YES NO
4. Generality probe
In another place and/or time, is it OK to do A? YES NO

The permissibility probe checks whether participants interpret the actions to be transgressions. The expectation is that participants will say Not-OK for all cases, indicating thereby that they indeed judge the actions to be instances of wrongdoing. The prediction related to both the authority contingency and generality probes is that participants will say Not-OK in cases of harmful transgressions (considering their wrongness to be authority independent and generalizable) and will say OK in cases of conventional transgressions (considering their wrongness cancelable by the decision of a legitimate authority and not necessarily extendable to different places and times). According to the Turiel tradition, these predictions have been borne out.

Contrary to Turiel's position, researchers have claimed that certain types of conventional transgressions evoke the moral signature. For example, Haidt, Koller, and Dias (1993) argued that conventional transgressions such as a person having sexual intercourse with a dead chicken or washing a toilet with the national flag evoke the moral signature. More recently, Kelly et al. (2007) claimed, conversely, that harmful transgressions do not generally evoke the moral signature. They argued that previous studies have shown that harmful transgressions evoke the moral signature because they have utilized only "schoolyard" transgressions (e.g., pulling hair). They suspected that if "grown-up" transgressions involving harm were studied, the relation between harm and authority independence or generality would collapse.

To test their suspicion, Kelly et al. created paired scenarios, each with a different "grown-up" harmful action, such as whipping as punishment, physical abuse as part of military training, and sleep deprivation as an interrogation method. In each pair, the harmful action remained constant and either the dimension of generality or the dimension of authority contingency was manipulated across scenarios. The following are examples of pairs related to the dimensions of authority contingency and generality:

WHIPPING AUTHORITY
(1) Mr. Adams is an officer on a large modern American cargo ship in 2004. One night, while at sea, he finds a sailor drunk at a time when the sailor should have been monitoring the radar screen. After the sailor sobers up, Adams punishes the sailor by giving him 5 lashes with a whip.
Question: Is it OK for Mr. Adams to whip the sailor? YES NO
(2) Now suppose that the Captain of the modern cargo ship had told Mr. Adams that 'On this ship it is OK for officers to whip sailors'.
Question: Is it OK for Mr. Adams to whip the sailor? YES NO
WHIPPING GENERALITY
(1) Mr. Adams is an officer on a large modern American cargo ship in 2004. One night, while at sea, he finds a sailor drunk at a time when the sailor should have been monitoring the radar screen. After the sailor sobers up, Adams punishes the sailor by giving him 5 lashes with a whip.
Question: Is it OK for Mr. Adams to whip the sailor? YES NO
(2) Three hundred years ago, whipping was a common practice in most navies and on cargo ships. There were no laws against it, and almost everyone thought that whipping was an appropriate way to discipline sailors who disobeyed orders or were drunk on duty.
Mr. Williams was an officer on a cargo ship 300 years ago. One night, while at sea, he found a sailor drunk at a time when the sailor should have been on watch. After the sailor sobered up, Williams punished the sailor by giving him 5 lashes with a whip.
Question: Is it OK for Mr. Williams to whip the sailor? YES NO

The first scenario of each pair presented the action in a way that, supposedly, participants would consider an instance of wrongdoing; hence, the OK question of the first scenario fulfilled the same role of the permissibility probe of the standard task. Each second scenario, together with its OK question, played a similar role to either the authority contingency probe or the generality probe of the standard task.

Each participant in Kelly et al.'s study responded to one of the paired scenarios. Kelly et al.'s prediction was that, for each of the paired scenarios, fewer participants would say OK to the first scenario question (the permissibility probe) than to the second scenario question (the authority contingency or generality related probes), a response pattern they consider to be incompatible with Turiel's hypothesis. They claimed that indeed their results confirm their prediction and disconfirm Turiel's hypothesis. For example, in relation to the whipping scenarios described above, they reported that, in the authority pair, only 6% said OK to the first scenario question, while 22% said OK to the second, and, in the generality pair, only 10% said OK to the first scenario question, while 51% said OK to the second.

We think that Kelly et al.'s perspective on the relationship between harm and morality has some problems, and we would like to develop another perspective, one more closely aligned with, but somewhat different from, the Turiel tradition.

First, Kelly et al. did not take into account the patterns of OK/Not-OK answers that could confirm or disconfirm

Turiel's hypothesis. Turiel's hypothesis concerning the moral side of the moral/conventional task is that participants who say Not-OK to the permissibility probe will also say Not-OK to the authority contingency and generality probes. To test Turiel's hypothesis, in each of Kelly et al.'s paired scenarios, one would have to compare the percentage of participants who answered Not-OK in the first scenario *and* Not-OK in the second (the pattern of responses confirming Turiel's hypothesis) with the percentage of participants who answered Not-OK in the first scenario *and* OK in the second (the pattern of responses disconfirming Turiel's hypothesis). Because Kelly et al. simply presented the pooled percentages of Yes answers to the first scenario and Yes answers to the corresponding second scenario, they did not report the response patterns that could test Turiel's hypothesis. Furthermore, it is important to notice that the responses of participants who answered OK in the first scenario are irrelevant to test Turiel's hypothesis—they cannot test whether *transgressions* involving a person being subject to harm evoke the moral signature, since they imply that the harmful action in question was not judged to be an instance of wrongdoing in the first place.

Second, we think that Turiel and associates' position is more nuanced than outlined by Kelly et al. The expectations that Turiel and associates would envisage for most of Kelly et al.'s "grown-up" harmful scenarios differ from the predictions they have advanced in relation to the harmful transgressions normally investigated using the moral/conventional task. In designing this task, Turiel and associates have deliberately utilized "prototypical transgressions" for the sake of investigating what is intrinsic to people's concepts about the moral domain. When talking about morality as related to a type of "domain" or "concern," the term "moral" is used with a different meaning, one more related to the content of norms, rather than to the moral signature. The Turiel tradition construes the moral domain as relating mainly to issues of harm, rights or justice and postulates an intrinsic relation between this domain and the moral signature (we return to this point later). In this paper, we use the expression "prototypical transgressions" to refer simply to the type of moral transgressions that have been *normally utilized in the moral/conventional task*—that is, "acts entailing harm, injustice, or violations of rights performed arbitrarily or for self-interested goals" (Wainryb, 1991, p. 842)—without implying that prototypical transgressions are more common in real life or that non-prototypical transgressions (those that do not fall into the prototypical category) are less common in real life. Most of Kelly et al.'s "grown-up" scenarios are not prototypical in this respect.

When judging whether a harmful action is an instance of wrongdoing, more is involved than the simple perception that the action has caused harm. Equally important is an understanding of the reasons that motivate the action—in particular, whether the reasons are considered invalid or valid. In many contexts, people do not judge a harmful action as an instance of wrongdoing because they view the reasons motivating the action as valid and therefore deem the causation of harm justified; for example, the harm inflicted in self-defense, in rites of initiation, in medical treatments, as punishment, or due to a commitment not to lie. Furthermore, in contexts such as these, there may be a fair amount of variation in people's judgments, both individually and cross-culturally, due to the fact that, in deciding whether a harmful action is justified, people may have divergent interpretations of the situation or may weigh differently the various factors involved, some not necessarily pertaining to the moral domain.

The Turiel tradition has discussed many of these non-prototypical cases, where perceptions of physical or psychological harm may be subordinated to (or coordinated with) other moral or non-moral considerations; for example, harm in the context of authoritative social pressures such as the Milgram experiments (Turiel & Smetana, 1984), in the context of cultural beliefs about the efficacy of harmful exorcism practices (Wainryb, 1993), and in the context of game rules permitting psychological harm (Helwig, Hildebrandt, & Turiel, 1995). The predictions of the Turiel tradition in relation to non-prototypical cases are much more qualified and, depending on each case, more or less specific.

In terms of the standard moral/conventional task, the Turiel tradition's broad expectation for non-prototypical cases would be that a significant percentage of participants would answer OK to the permissibility probe. Furthermore, concerning the rationales guiding participants' overall patterns of response to the permissibility and moral signature probes, the broad expectation would be that considerations not directly (and not necessarily) related to moral concerns would play an important role.

A study by Wainryb (1991) on judgments of wrongdoing related to corporal punishment illustrates this point. Wainryb compared participants' judgments of a prototypical harmful transgression in the sense described above (a father who out of frustration spanks his son who has done nothing wrong) with judgments of a non-prototypical one (a father who spanks his son for repeatedly misbehaving). In relation to the prototypical transgression, she found that all participants said Not-OK to both the permissibility and authority contingency probes. Furthermore, 96% of participants justified their answers in terms of welfare concerns and the fact that the harm was unprovoked. As predicted, the results were quite different in response to the non-prototypical harmful action. First, there was substantial variation in the evaluations related to the permissibility probe (57% said OK, 43% said Not-OK). Second, about 94% of participants' explanations for their answers involved factual beliefs (what Wainryb calls "informational assumptions") about the efficacy of spanking as a teaching method—those who answered OK considered spanking effective, those who answered Not-OK considered spanking ineffective or of uncertain efficacy.

Wainryb manipulated two other probes in relation to the non-prototypical harmful action. One was the authority contingency probe; depending on the participant's original answer to the permissibility probe, the opposite authority contingency information was posited: "Suppose that respected leaders in the community said that it is wrong/alright for a father to spank his child." The other probe presented participants with opposite expert opinion on the efficacy of spanking and asked them to hypothetically assume its correctness. For example, for a participant who originally assumed the efficacy of spanking: "Suppose that experts who know a lot about the ways children learn could prove that spanking does not teach children anything and that children do not learn when they get spanked. If you were convinced by what they say and you believed that it was true, would you then think it was alright or not alright for a parent to spank his child?" Wainryb found that whereas in the authority contingency probe the great majority of participants maintained their original OK or Not-OK answer to the permissibility probe, in the expert probe the great majority of participants changed their original evaluation. In other words, the normative force of respected leaders was not enough to change participants' original judgments, but changes in informational assumptions based on expert opinion were. In sum, participants' different patterns of response in Wainryb's study concerning the non-prototypical harmful action were driven by diverging assumptions about the utility of spanking as a pedagogical tool—an assumption not directly (and not necessarily) related to the moral domain. (Notice that we are not claiming that utilitarian reasoning in general or the utilitarian reasoning related to this non-prototypical harmful action in particular are necessarily unrelated to moral concerns—thus, we are not claiming that non-prototypical harmful actions are necessarily unrelated to moral concerns either. We are simply claiming that the assumption about efficacy, in itself, is not a moral concern and is not necessarily related to the moral domain.)

A third problem with Kelly et al.'s perspective is that their claims and arguments were based only on participants' OK or Not-OK answers. One general lesson from our discussion of non-prototypical scenarios (in particular, from Wainryb's study) is that one should not take participants' patterns of OK/Not-OK responses to these cases at face value: even patterns that apparently confirm or disconfirm Turiel's hypothesis may be, in actuality, irrelevant to testing it because they *may* be driven by rationales that are not related to the moral domain. Kelly et al.'s study did not include any justification probe; thus, it does not offer any information on participants' rationales for their OK or Not-OK answers.

Given the fundamental importance of providing some evidence on participants' rationales, in our replication of Kelly et al.'s "grown-up" scenarios, we included a justification probe asking participants to explain each of their OK/Not-OK answers. We relied heavily on the analysis of this qualitative data to test the general hypothesis that harmful transgressions evoke the moral signature. However, as it stands, this general hypothesis is vague. In order to state our prediction more precisely, we need to specify the concept of harmful transgression that is at stake. One has to be explicit about what is supposed to establish that a harmful action is a transgression and that its wrongness is authority independent and general in scope. One cannot simply say that harmful transgressions evoke the moral signature because they are perceived to cause harm, since harmful actions in themselves need not even be considered transgressions.

Kelly et al. (2007) interpreted Turiel's hypothesis as follows: "Transgressions involving harm, justice, *or* rights evoke the signature moral pattern" (p. 120; our emphasis). It is as if there were three separate types of moral transgressions (those involving harm, those involving injustice, and those involving rights violation) that independently evoke the moral signature. This characterization does not clarify what is supposed to establish that harmful actions are transgressions and evoke the moral signature, although it seems to imply that harmful transgressions do not necessarily have to be tied to perceptions of injustice and rights violation in order to evoke the moral signature. For this reason, when Kelly et al. proposed, *contra* Turiel, that harm does not generally evoke the moral signature (i.e., harm is not sufficient for morality), either they were claiming that harmful transgressions do not evoke the moral signature without specifying the type of transgression at stake (i.e., what leads harmful actions to be perceived as transgressions in the relevant cases) or they were claiming that not all types of harmful actions evoke the moral signature, which is true, but simply because the causation of harm in itself does not need to be perceived as a transgression.

Turiel and associates indeed frequently talk about harm as connected to one of three distinct moral dimensions. For example, in Wainryb's (1991) study sketched above, she included three distinct types of prototypical moral transgressions (related to welfare, justice, or rights):

> The study included ... acts entailing harm, injustice, or violation of rights performed arbitrarily or for self-interested goals. This type of event, hereafter referred to as *prototypical moral violations* (PM), allowed for the assessment of concepts of welfare, justice, and rights when they are not in conflict with other considerations ... Welfare (W). – The prototypical moral violation (PM-W) described a father who out of frustration spanked his son who has done nothing wrong. . .Justice (J). – The prototypical moral violation (PM-J) described a store manager who refused to interview qualified women for a job because he did not like women. . .-Rights (R). – The prototypical moral violation (PM-R) depicted a mayor who made all Chinese children in town attend a separate school because he did not like them to mingle with non-Chinese children. (pp. 842–843)

They also seem to accept that the three dimensions somewhat independently evoke the moral signature, with the implication that perceptions of injustice and rights violation are *orthogonal* to the characterization of harmful transgressions that evoke the moral signature. However, they suppose that what establishes harmful actions as transgressions evoking the moral signature is that they are tied to perceptions of welfare violations.

Here our perspective departs from the way Turiel and associates frame the discussion of the relation between harm and morality. In our view, perceptions of injustice or rights violation are fundamental to the proper characterization of the hypothesis *harmful transgressions evoke the moral signature*. Take the prototypical example of harmful transgression quoted above—the father who, out of frustration, spanked his son who had done nothing wrong. We do not think that people's evaluations of this case are disconnected from perceptions of injustice or rights violation—the son does not *deserve* to be treated this way; it goes against his basic rights.[1] More generally, we believe that the prototypical harmful transgressions that have been utilized in the moral/conventional task similarly involve perceptions of injustice or rights violation. Thus, we reinterpret Turiel's hypothesis as follows: transgressions involving harm *and* injustice or rights violation evoke the moral signature, or, more explicitly, harmful transgressions are conceived to be authority independent and general in scope if they are perceived to entail injustice or rights violations. Notice also that, with this characterization, we are not denying that there are other types of *transgressions* that evoke the moral signature. Likewise, we are not denying that there are other normative domains that evoke the moral signature (*cf.* Haidt et al., 1993).

We can now delineate our specific prediction concerning the replication of Kelly et al.'s study. Our prediction on the relation between harm and the moral signature was the following:

> Whenever a participant answers Not-OK to the permissibility probe *and* their answer is driven by concerns with justice or rights (which include welfare), the answer to the moral signature (authority contingency or generality) probe will be Not-OK as well based upon the same concerns.
>
> The alternative hypothesis predicts the following:
>
> Even if a participant answers Not-OK to the permissibility probe and their answer is driven by concerns with justice or rights (which include welfare), the answer to the moral signature (authority contingency or generality) probe may be OK based upon the normative force of an authority or social consensus that is distant in space or time.

Furthermore, in accordance with Turiel and associates' general position on non-prototypical harmful scenarios, we have the following broad expectations:

> Some responses to the permissibility and moral signature probes will be guided by concerns not directly (and not necessarily) related to the moral domain; also, some participants will perceive the harmful actions in question as justified even when they are not supported by present-day authority or social consensus, and, consequently, will answer OK to the permissibility probe.

2. Method

2.1. Participants

As in Kelly et al.'s study, participants were recruited via links to websites for online psychological research ("Psychological Research on the Net" [http://psych.hanover.edu/research/exponnet.html]; "Social Psychology Network" [http://www.socialpsychology.org/expts.htm]) between the period of April 23rd and June 7th, 2007. Links to the online survey were also emailed to undergraduate students from Queen's University, Belfast. Participation was anonymous, without compensation, and restricted to those 18 years of age and over. This generated a total sample of 159 adult participants who answered all parts of the survey (53% female, 47% male), of whom 12 identified themselves as living outside of the United States.

2.2. Materials and procedure

Kelly et al.'s study included eight paired scenarios involving harmful actions. In the replication, we utilized five of their "grown-up" pairs. In addition to the two whipping pairs described previously, the following three pairs were included in our study:

> PRISONER AUTHORITY
> (1) Sergeant Johnson is interrogating a suspected terrorist who may have information about future terrorist attacks. His commanding officer has ordered him not to use sleep deprivation as a way of getting information. Nonetheless Sergeant Johnson keeps the suspect awake for three days and three nights.
> *Permissibility Probe*: Is it OK for Sergeant Johnson to keep the suspect awake for three days and three nights?
> YES NO
> *Justification Probe:* Please thoroughly explain why you marked this option.
> (2) Now suppose that before he decided to keep the prisoner awake, Sergeant Johnson's commanding officer had told him that the use of sleep deprivation is an acceptable way of trying to get information when interrogating suspected terrorists, and that Sergeant Johnson could use sleep deprivation whenever he wanted to.
> *Authority-related Probe*: Is it OK for Sergeant Johnson to keep the suspect awake for three days and three nights?
> YES NO
> *Justification Probe:* Please thoroughly explain why you marked this option.
> TRAINING AUTHORITY
> (1) For many years, the military training of elite American commandos included a simulated interrogation by enemy forces in which the trainees were threatened and physically abused. Most people in the military believe that these simulated interrogations were helpful in preparing trainees for situations they might face later in their military careers. Though no one was ever killed or permanently disabled by the physical abuse they received during these simulated interrogations,

[1] As we mentioned earlier, in coding participants' justifications for this transgression, Wainryb claimed that 96% were related to the categories *welfare* and *unprovoked harm*. But we do not think that questions of welfare are disconnected from questions of rights in a broad sense, nor do we think that notions of unprovoked harm are disconnected from questions of desert. For this reason, when dealing with participants' justifications in what follows, we shall a adopt a more global way of coding the data, instead of following coding schemes that have been utilized by the Turiel tradition, which parse the moral domain into different dimensions (see Davidson, Turiel, & Black, 1983, for an influential formulation).

the trainees often ended up with bruises or injuries that lasted for a week or more.

Recently, the Pentagon issued orders prohibiting physical abuse in military training. Sergeant Anderson is a soldier who trains elite American commandos. He knows about the orders prohibiting physical abuse and his immediate superiors have ordered him not to do it. Nonetheless, he regularly threatens and physically abuses trainees during the simulated interrogations that he conducts.

Permissibility Probe: Is it OK for Sergeant Anderson to physically abuse trainees during simulated interrogations? YES NO

Justification Probe: Please thoroughly explain why you marked this option.

(2) Now suppose that the Pentagon had never issued orders prohibiting physical abuse in military training, and that Sergeant Anderson's superiors had told him that the use of physical abuse was acceptable in simulated interrogations.

Authority-related Probe: Is it OK for Sergeant Anderson to physically abuse trainees during simulated interrogations? YES NO

Justification Probe: Please thoroughly explain why you marked this option.

SLAVERY GENERALITY

(1) In the United States, slaves were an important part of the economy of the South 200 years ago. American slaves were used mainly to maintain households, and to supply agricultural labor.

Permissibility Probe: Was it OK for Americans to keep slaves? YES NO

Justification Probe: Please thoroughly explain why you marked this option.

(2) In ancient Greece and Rome, slaves were an important part of the economic and social system. Greek and Roman slaves were used as oarsmen, as soldiers, to maintain households, and to supply agricultural labor.

Generality-related Probe: Was it OK for the ancient Greeks and Romans to keep slaves? YES NO

Justification Probe: Please thoroughly explain why you marked this option.

As in Kelly et al.'s study, the order of presentation of each pair of scenarios was counter-balanced, and each participant was randomly assigned to one of the five paired scenarios in one of its two possible orders (i.e., either with the first scenario of the pair, the one related to the permissibility probe, presented initially or not). This was one way in which Kelly et al.'s procedure differed from the standard task—in the latter, the order of presentation is normally fixed, with the permissibility probe and its related scenario presented initially. In this paper, we use the expression "first scenario" to refer to the scenario related to the permissibility probe, and the expression "second scenario" to refer to the scenarios related to the moral signature (authority or generality) probes.

Participation took place online via a website titled 'Five Minute Morality Survey.' The online survey was designed using SurveyMonkey© survey builder. Before answering the survey, participants gave their informed consent. Next, participants were directed to a page with the first or second scenario and the corresponding probes. The first probe asked participants whether it was OK for the protagonist of the scenario to engage in the action described, with the possibility of a Yes or No answer. The second probe asked participants to justify their Yes or No answer. On the next page, participants were presented the other scenario and the two corresponding probes. After responding to the paired scenarios, participants completed a brief demographic questionnaire, and then read a debriefing statement, which explained the purpose of the study and thanked them for participating.

2.3. Coding and reliability

In each paired scenario, we delineated the patterns of Yes (OK) or No (Not-OK) answers that could test our hypothesis. We coded the percentage of No–No, No–Yes, Yes–Yes, and Yes–No response patterns (the first Yes or No are related to the first scenario; the second Yes or No to the second scenario). Yes–Yes and Yes–No patterns are irrelevant to test our hypothesis, because they indicate that the participant did not judge the harmful action as a transgression in the first place. A No–No pattern is *prima*

Table 1
Justification categories and definitions.

Category	Definition
Restatement	Justification simply appeals to the acceptability *or* unacceptability of the action without providing any additional scorable justification
Justice/rights/welfare	Justification appeals to justice/rights/welfare *or* to violation of justice/rights/welfare
Utility	Justification appeals to the usefulness *or* non-usefulness of the action in achieving an intended purpose
Social norms	Justification appeals to whether *or* not the action follows the socially established norms without indicating the reason why one should follow *or* should not violate the social norms. If it indicates this underlying reason, the justification is coded according to this reason
Authority	Justification appeals to whether *or* not the action obeys an authoritative command without indicating the reason why one should obey *or* not obey the authoritative command. If it indicates this underlying reason, the justification is coded according to this reason
Personal conscience	Justification appeals to whether *or* not the action follows the agent's own sense of right and wrong
Informed consent	Justification appeals to whether *or* not the recipient of harm had prior knowledge about and had consented to the possibility of being injured or punished
Unscorable	No justification; justification unclear; justification does not fall into any of the preceding categories

facie evidence confirming our hypothesis and a No–Yes pattern is *prima facie* evidence disconfirming it. These patterns constitute no more than *prima facie* evidence because our hypothesis cannot be tested independent of evidence on the types of rationales guiding the Yes/No answers.

Driven by our theoretical interests and the patterns apparent in the qualitative data, we developed a coding scheme to analyze each explanation given in response to the justification probe (see Table 1).

The first thing to notice in this coding scheme is that each of the categories may apply to justifications of either Yes or No answers. A participant might justify an OK evaluation of whipping a sailor caught drunk on watch by saying, "it is a fair punishment." Conversely, a participant might justify a Not-OK evaluation to whip the same sailor by saying, "it is against human rights." Each of these explanations would be coded as *justice/rights/welfare* (*JRW*). Related to the category *utility*, a participant might justify an OK evaluation of sleep deprivation by saying, "it is an effective means of extracting reliable information." Conversely, a participant might justify a Not-OK evaluation of the same action by saying, "it is not an effective means of extracting reliable information."

The categories *social norms* and *authority* include a clause that was introduced to isolate the underlying rationale guiding participant's OK or Not-OK evaluations. Many participants referenced social norms or authority in their explanations, but directly qualified these references in terms of the reasons for having the specific social norms or order. These cases were coded according to the reasons motivating the social norm or order. For example, a participant might justify an OK evaluation of trainee abuse by saying, "the Pentagon allows the abuse of trainees in order to guarantee successful training." In this case, the explanation would be coded as *utility*, rather than *authority*. A participant might justify a Not-OK evaluation of slavery by saying, "it is a violation of societal rules protecting human rights." In this case, the explanation would be coded as *JRW*, rather than *social norms*. In the same fashion, any other coding category could, in principle, constitute an underlying reason. However, it is important to notice that our coding scheme does not preclude a participant providing a justification belonging to the category *social norms* or *authority* in addition to justifications from a different category. For example, a participant might justify a Not-OK evaluation of whipping a sailor by saying, "it is wrong to disobey orders *and* this violence is cruel." In this case, each justification would be coded independently; that is, the explanation would be coded as including both an instance of *authority* and an instance of *JRW*.

The general point of introducing the clause in relation to *authority* and *social norms* is to discern when a reference to authority or social norms constitutes a distinct form of justification. We think this is the case when the reference indicates that the OK or Not-OK evaluation of the action is based upon an acceptance of the normative force of the authority or social consensus *in and of itself* (i.e., independent of the normative content promoted by sanctions or social norms).

When a participant provided multiple justifications for their Yes or No answer, we handled them as follows. If the justifications were instances of different categories, they were coded independently (as in the last example). If the justifications were different instances of the same category (i.e., the explanation was redundant), they were coded as one instance of that category. For example, a participant might justify a Not-OK evaluation of whipping a sailor by saying, "it is against human rights, unfair, cruel, and inhumane." In this case, these justifications would be coded together as one instance of the category *JRW*.

Participants' explanations for the two scenarios were coded independently. In other words, the explanation given by a participant in relation to the first scenario was not taken into account when coding the explanation given to the second scenario (and vice versa). The only exception was when participants explicitly directed the researchers to their previous explanation, as when they stated, "see previous answer." In this case, the coding of the previous explanation was used.

The authors independently coded all explanations and together reached agreement on the overall coding of the data. Then, an independent coder unfamiliar with moral psychology and the purposes of the study coded all explanations. The inter-rater agreement between the authors and the independent rater using the coding scheme was more than satisfactory (Cohen's kappa = .80).[2]

In terms of our coding categories, we shall now characterize what constitutes *real* evidence, as opposed to simply *prima facie* evidence, to test our hypothesis:

> *Weak* evidence confirming our hypothesis consists of cases in which a No answer to the permissibility probe is justified in terms of instances of *JRW* and the overall pattern of response is No–No. *Strong* evidence consists of cases in which, more specifically, the second No answer is justified in terms of instances of *JRW* too.
>
> *Weak* evidence disconfirming our hypothesis consists of cases in which a No answer to the permissibility probe is justified in terms of instances of *JRW* and the overall pattern of response is No–Yes. *Strong* evidence consists of cases in which, more specifically, the Yes answer is justified in terms of instances of *authority* or *social norms*.

3. Results

The percentage of participants evincing the No–No, No–Yes, Yes–Yes and Yes-–No response patterns in each of the paired scenarios of the replication and of Kelly et al.'s study is shown in Table 2.[3]

[2] For access to a more detailed explanation of the coding scheme, including the coding instructions and the dummy data used to train the independent coder, contact one of the authors.

[3] Kelly et al. provided us with the percentages not reported in their original article. Overall, the Yes–No response pattern was evinced only by three participants in their study (each in a different paired scenario) and one participant in the replication. This pattern is completely counterintuitive in the context of these paired scenarios and the justifications given by our participant do not seem serious. For this reason, we eliminated these "outliers" from the analysis.

Table 2
Percentage of participants for each response pattern.

Scenarios	Study	N	No–No (%)	No–Yes (%)	Yes–Yes (%)	Yes–No (%)
Slavery Generality	Replication	30	97.0	0.0	3.0	0.0
	Kelly et al.	187	88.0	5.0	7.0	0.0
Whipping Generality	Replication	30	77.0	16.0	7.0	0.0
	Kelly et al.	198	49.0	41.0	10.0	0.0
Whipping Authority	Replication	33	94.0	3.0	3.0	0.0
	Kelly et al.	195	76.5	17.5	6.0	0.0
Prisoner Authority	Replication	34	79.0	18.0	3.0	0.0
	Kelly et al.	172	85.0	14.5	0.5	0.0
Training Authority	Replication	31	36.0	48.0	16.0	0.0
	Kelly et al.	150	42.0	49.0	9.0	0.0

It is important to notice that (i) except for Whipping Generality, the results of the replication are quite similar to Kelly et al.'s results, (ii) except for Whipping Generality (Kelly et al.'s results) and Training Authority (both results), the percentages of the No–No response pattern are fairly high, and (iii) overall, a non-trivial percentage of participants evinced the Yes–Yes response pattern.

The pooled number of justifications by category for each Yes or No answer (separated by response pattern), for each paired scenarios, is shown in Table 3. To illustrate how the table is formatted, notice that, for the No–No response pattern related to Slavery Generality, there are four instances of *restatement* justifying No answers to the permissibility probe, and four instances of *restatement* justifying No answers to the generality probe. However, since these numbers were pooled separately, one cannot infer that instances of *restatement* justifying the first No and instances of *restatement* justifying the second No come from the same participants; for example, a participant who justified the first No in terms of *restatement* may have justified the second No in terms of *JRW*, or a participant who justified the first No in terms of *JRW* may have justified the second No in terms of *restatement*.

It is important to notice that (i) except for Training Authority, the number of *JRW* justifications was extremely large for all paired scenarios, (ii) *JRW* justifications were mostly related to answers of the No–No response pattern, and (iii) except for Slavery Generality and especially in Training Authority, participants utilized a variety of rationales to justify their Yes or No answers.

To allow a more focused understanding of the way (and extent to which) participants appealed to *JRW* justifications in those response patterns that could test our hypothesis, we divided participants into three groups—those with *JRW* justifications in both, one, or none of the paired scenarios (see Table 4).

It is important to notice that (i) in the No–No response pattern, only a small number of participants did not use *JRW* justifications (i.e., were in the category "None"), (ii) in the No–Yes response pattern, about half of our participants did not use *JRW* justifications, and the majority of these were from Training Authority, and (iii) in the No–No response pattern, the great majority of participants who used *JRW* justifications, used them in response to both scenarios.

4. Discussion

Kelly et al. (2007) presented findings that allegedly call into question a well-established finding of the moral psychology literature—that harmful transgressions are considered to be authority independent and general in scope (e.g., Turiel, 1983). We advocated a particular construal of this hypothesis and replicated Kelly et al.'s study, modifying certain methodological parameters regarding research design and data analysis. The results provided strong confirmation for our hypothesis and were consistent with our broad expectations.

The results gave *prima facie* confirmation for our hypothesis. It is important to notice that the percentages of No–No and No–Yes answers described in Table 2 are not entirely precise. Given that the Yes–Yes response pattern is irrelevant to test our hypothesis, one would have to recalculate the percentages excluding participants with this pattern from the total. For example, the revised No–No and No–Yes percentages for Slavery Generality would be, respectively, 100% and 0% in the replication (94% and 6% in Kelly et al.' s study). When the irrelevant cases are removed, it becomes still more apparent that there is greater *prima facie* evidence confirming our hypothesis than disconfirming it, although in Whipping Generality (in Kelly et al.'s study) and Training Authority (in both studies) the balance between confirmation and disconfirmation remains roughly even.

The results of participants' explanations for No–No answers indicated that much of this *prima facie* evidence is not only *real* evidence, but also *strong* evidence confirming our hypothesis. The great majority of No–No participants utilized *JRW* justifications in their explanations, which is a necessary condition for *prima facie* evidence to constitute real evidence. These participants offered explanations such as[4]:

(i) 1(No): "Slavery is an unacceptable violation of someone's free will and human rights in general" 2(No): "...violation of human rights" [Slavery Generality; S; 1: *JRW*; 2: *JRW*]

[4] Below, 'S' stands for standard order of presentation, that is, the permissibility probe being presented initially ('NS' for non-standard order), '1' indicates first scenario ('2' the second), and '+' indicates that multiple justifications were coded independently.

Table 3
Pooled number of justifications by category for each Yes or No answer (separated by response pattern).

Scenarios	Justification category	No	No	No	Yes	Yes	Yes
Slavery Generality	Restatements	4	4	–	–	–	–
	Justice/rights/welfare	24	22	–	–	–	–
	Utility	–	–	–	–	1	–
	Social norms	–	–	–	–	1	1
	Authority	–	–	–	–	–	–
	Personal conscience	–	–	–	–	–	–
	Informed consent	–	–	–	–	–	–
	Unscorable	1	3	–	–	–	–
Whipping Generality	Restatements	1	5	–	–	–	–
	Justice/rights/welfare	16	16	3	–	–	1
	Utility	3	2	1	2	–	–
	Social norms	3	1	2	5	1	1
	Authority	–	–	–	–	–	–
	Personal conscience	–	–	–	–	–	–
	Informed consent	–	–	1	1	1	1
	Unscorable	3	1	–	–	1	1
Whipping Authority	Restatements	3	2	–	–	–	–
	Justice/rights/welfare	20	21	1	–	1	–
	Utility	3	3	–	–	1	–
	Social norms	2	1	–	–	–	–
	Authority	2	2	–	–	–	–
	Personal conscience	–	–	–	–	–	1
	Informed consent	–	2	–	1	–	–
	Unscorable	2	2	–	–	–	–
Prisoner Authority	Restatements	1	2	–	–	–	–
	Justice/rights/welfare	20	20	3	4	1	1
	Utility	9	8	3	4	1	1
	Social norms	1	1	–	–	–	–
	Authority	12	–	1	1	–	–
	Personal conscience	–	–	–	1	–	–
	Informed consent	–	–	–	–	–	–
	Unscorable	–	2	–	1	–	–
Training Authority	Restatements	3	1	–	–	–	–
	Justice/Rights/Welfare	5	8	2	3	2	2
	Utility	–	2	1	8	5	5
	Social norms	1	–	2	–	–	–
	Authority	6	–	10	–	–	–
	Personal conscience	–	–	–	1	1	–
	Informed consent	–	–	1	7	1	2
	Unscorable	–	–	1	2	–	–

Table 4
Number of participants with Justice/rights/welfare justifications in both, one (first or second), or none of the paired scenarios.

Response pattern	Scenarios	*N*	Both	1st	2nd	None
No–No	Slavery Generality	29	21	3	1	4
	Whipping Generality	23	14	2	2	5
	Whipping Authority	31	17	3	4	7
	Prisoner Authority	27	17	3	3	4
	Training Authority	11	5	–	3	3
No–Yes	Slavery Generality	0	–	–	–	–
	Whipping Generality	5	–	3	–	2
	Whipping Authority	1	–	1	–	–
	Prisoner Authority	6	2	1	2	1
	Training Authority	15	–	2	3	10

(ii) 1(No): "It takes away one of the sailor's basic human rights, i.e., the right not to be physically hurt" 2(No): "Just because something is accepted as common practice does not therefore make it right" [Whipping Generality; S; 1: *JRW*; 2: *restatement*]

(iii) 1(No): "...this is against their human rights, it's abuse" 2(No): "Just because its the way the ship works does not mean it is right to whip a sailor" [Whipping Authority; S; 1: *JRW*; 2: *restatement*]

(iv) 1(No): "A soldier is supposed to follow orders and his actions where inhumane" 2(No): "Even though he was following orders, that doesn't excuse human rights abuse" [Prisoner Authority; S; 1: *authority + JRW*; 2: *JRW*]

(v) 1(No): "It is not ok to hurt someone, and if you have been told not to, this makes the deed even more bad" 2(No): "it is not ok to hurt someone, not even in the name of 'training-for-combat', and especially not when someone in a position of power abuses that position by hurting subordinates who cannot object to it" [Training Authority; NS; 1: *JRW + authority*; 2: *JRW*]

Furthermore, most No–No participants used *JRW* justifications in responding to both scenarios, indicating that their answers constitute strong confirmation for our hypothesis. We should also add that, as some of the examples above show, participants who used *JRW* justifications only once appealed frequently (10 out of 24 times) to instances of *restatements* in responding to the other scenario, which may be simply an economical way of repeating the previous type of justification. Thus, it is not too much speculation to say that many of the No–No answers that constitute only weak evidence were in fact guided by "strong" rationales.

On the other hand, the fact that about half of participants' explanations for No–Yes answers do not involve *JRW* justifications shows that a substantial portion of the *prima facie* disconfirmation does not constitute real evidence. Moreover, when one inspects the explanations of the other half, the No–Yes results are much less straightforward than the No–No results.[5] Consider the following explanations:

(i) 1(No): "Modern times uses more effective means of punishment. Whipping is not expected nor deserved in this situation" 2(Yes): "Because it was a generally accepted means of punishment that was most likely understood by the sailor" [Whipping Generality; NS; 1: *utility + informed consent + JRW*; 2: *social norms + informed consent*]

(ii) 1(No): "No one has the right to punish another person by inflicting pain..." 2(Yes): "As long as [the sailor] understood the circumstances in which it may be possible where he would be whipped. He should have a good understanding of the rules and regulations and so should know that not adhering to such would result in punishment" [Whipping Authority; S; 1: *JRW*; 2: *informed consent*]

In the first example, because we coded the second explanation as involving an instance of *social norms* in addition to an instance of *informed consent*, the No–Yes answer of the participant constitutes strong evidence against our hypothesis. However, we are not sure that this is the best interpretation. It is not clear that the participant is saying that the social consensus in itself constitutes a normative force with the power to cancel the wrongness of whipping. As an alternative interpretation, on might say that the rationale guiding the Yes answer is simply related to the fact that, when the sailor took the job, he was aware of (and accepted) that one could be punished with whipping, in which case the explanation would be coded simply as involving an instance of *informed consent*. The second example illustrates the possibility of this alternative interpretation (there is not much difference between Whipping Authority and Whipping Generality regarding the aspect at issue here). Consider a few more examples:

(iii) 1(No): "Johnson has no reason to disregard an order. By doing so, he is being cruel to the suspect" 2(Yes): "Sergeant Johnson is following orders. In order for him to be a soldier, he has to learn to trust those in command above him" [Prisoner Authority; NS; 1: *JRW*; 2: *authority*]

(iv) 1(No): "If ordered not to, there is no reason to disobey as this person could be innocent or for a myriad of other reasons besides the fact that it's inhumane. Torture doesn't guarantee information either" 2(Yes): "If there is sufficient evidence to prove they might be suspect, then yes, as a one time deal. I don't believe it's right to detain and torture repeatedly "suspected terrorists" if there isn't any real evidence other than circumstantial or no new information is being brought forth" [Prisoner Authority; NS; 1: *JRW + utility*; 2: *JRW + utility*]

(v) 1(No): "If there had been a direct order to not use the technique, there must be a good reason to explicitly not allow sleep deprivation. Also, without the order, any incentive to use sleep deprivation is gone as the suspect may not have important enough information to be subjected to it" 2(Yes): "Given the situation involving a suspected terrorist, it is acceptable to use this non violent technique to obtain information" [Prisoner Authority; NS; 1: *utility*; 2: *JRW + utility*].

In the third example, because we coded the second explanation simply as involving an instance of *authority*, the No–Yes answer of the participant constitutes strong evidence against our hypothesis. However, again, we are not sure that this is the best interpretation. It is not clear here that the participant is saying that the authority in itself constitutes a normative force with the power to cancel the wrongness of sleep deprivation. One might say that it all depends on the type of "trust" the participant has in mind when they write, "he has to learn to trust those in command above him." One alternative interpretation is that the trust at stake here is simply related to the accuracy of the factual knowledge possessed by the superiors regarding whether the prisoner is a real suspect or not, whether the suspect has the relevant information or not, or whether the method of using sleep deprivation is effective or not. In this type of interpretation, as in Wainryb's (1991) study (discussed in the introduction), participants

[5] There is an additional serious problem with the No-Yes results that we shall not discuss here. In brief, the problem is related to a certain type of polysemy of the OK questions of the standard moral/conventional task that is particularly detrimental in the context of Kelly et al.'s (and our) design. For a discussion, see Sousa, 2009.

are changing their evaluations of the use of sleep deprivation because they accept that superiors possess "expert" knowledge about the efficacy of using sleep deprivation for obtaining relevant information, including information that may potentially be used to prevent future harm, and therefore safeguard the general welfare. Examples (iv) and (v) illustrate the possibility of this alternative interpretation.

The results of our replication were also consistent with our broad expectation that in non-prototypical harmful cases some participants would be guided by concerns not directly (and not necessarily) related to the moral domain. These types of concerns are even more pronounced in participants' explanations of No–Yes answers related to Training Authority. This condition provided the greater *prima facie* evidence against our hypothesis, although most of it did not fulfill the criteria for real evidence (see Table 4). Participants offered explanations such as:

(i) 1(No): "...it is against policy and he was directed not to do it" 2(Yes): "If it is part of the training and the candidates know it is going to happen prior to the event, then it is acceptable..." [Training Authority; NS; 1: *social norms* + *authority*; 2: *informed consent*]

(ii) 1(No): "If the superiors are saying he should not do it, this implied to me that it is not a codified part of the training program and this trainees have not agreed to it" 2(Yes): "I am assuming that the trainees have given "informed consent" to the training, including the possible physical abuse, and are of sound mind and body. if not, then I would say it is not OK for him to abuse them" [Training Authority; NS; 1: *informed consent*; 2: *informed consent*]

(iii) 1(No): "Abuse is no longer part of the mandated training program. Employing abuse would go against the orders given to Anderson" 2(Yes): "The abuse occurs in the confines of a training exercise designed to prepare soldiers for the type of treatment they may receive ... the soldiers are aware that such training tactics are employed and therefore, give consent to the abuse" [Training Authority; NS; 1: *authority*; 2: *utility* + *informed consent*]

(iv) 1(No): "Although Sergeant Anderson might believe what he is doing is for the greater good and that his abuse rears stronger commandos, if he has an order from superiors, he should follow it" 2(Yes): "The abuse that a commando is subjected to prepares him or her for what is to come in the future..." [Training authority; S; 1: *authority*; 2: *utility*]

(v) 1(No): "The same rationale applies: if those fully informed in the possible value of a training technique decide it shouldn't be used, ethically the Sgt. should desist" 2(Yes): "I personally doubt that superiors' permission alone is sufficient to morally justify such extreme behavior. However, the consensus within the military community (i.e., the individuals most qualified to judge the usefulness of the "abuse" as a training tool) are the only ones who are in a position to judge the ethics of the situation" [Training Authority; NS; 1: *utility*; 2: *utility*]

The first three examples demonstrate a different subtype of *informed consent*. In this case, the rationale is related to harm being envisaged and accepted in the context of participating in an established training procedure, instead of relating to the recipient of harm envisaging and accepting the possibility of being harmed when doing something wrong. The last three examples show *precautionary reasoning* (see Fiddick, 2004; *cf.* Tisak & Turiel, 1984) concerning the benefits of inflicting harm (for the soldier and the military system, and perhaps also for the general public, given the protective role of the military), based on informational assumptions about the efficacy of the training procedure. The final example demonstrates that these assumptions may be calibrated by trust in expert knowledge.

The results of our replication were consistent with our other broad expectation that in non-prototypical cases some participants would perceive the harmful actions as justified even when they are not supported by present-day authority or social consensus. Participants' explanations for Yes–Yes answers illustrate our expectation and highlight some additional points as well:

(i) 1(Yes): "Sure, if you are dumb enough to sign up for the Navy during a war, you are lucky to only come out with 5 lashes from a whip! Seriously though, I think the punishment should be whatever is socially acceptable AND must be known to all. Misbehaviors and consequences should be defined before either arises. Then if you decide to break the (known) rules, you will receive the (known) punishment" 2(Yes): "...it was the socially acceptable punishment at the time. The sailor knew what the consequences for drunken behavior were and decided to get drunk anyway. His drunkeness could have jeopardized the safety of the ship. If anything the sailor should have gotten more lashes! [Whipping Generality; NS; 1: *social norms* + *informed consent*; 2: *social norms* + *informed consent* + *JRW*]

(ii) 1(Yes): "He won't get drunk again. He could have risked the lives of his sea mates" 2(Yes): "Sometimes you must judge how to address a situation yourself..." [Whipping Authority; NS; 1: *utility* + *JRW*; 2: *agent conscience*]

(iii) 1(Yes): "...the abuse is an important part of this particular type of training. The government has to say no to it because it may make them look bad in the public eye..." 2(Yes): "...it is always ok for Sergeant Anderson to abuse the recruits because I see it as a vital part of their training" [Training Authority; S; 1: *utility*; 2: *utility*]

(iv) 1(Yes): "I think if the man is a terrorist the information will save many people's lives – if he's not I think this will come out and sleep deprivation is not necessarily fatal..." 2(Yes): "see previous answer" [Prisoner Authority; S; 1: *utility* + *JRW*; 2: *utility* + *JRW*]

In the first example, the participant misinterpreted the scenario as being related to a war context where whipping as punishment would be expected. However, leaving aside this problem, the rationale invoked across explanations reveals that the presence or absence of informed consent is being considered fundamental to the judgment of whether or not the manner of punishment is justified. This point may have been implicit in other participants' explanations that utilized instances of *informed consent* in the context of punishment.[6] In the first and second examples, the culpability of the sailor, given his negligent behavior, was evoked in justifying (and even in prescribing more) punishment. The second example also raises the possibility of a case where the harmful action is being judged to be *morally right*, since the deserved punishment seems to be considered independent of authority (i.e., seems to evoke the moral signature). In the second, third, and fourth examples, the utility of the harmful actions (i.e., whipping as a deterrent, sleep deprivation as providing important information, physical abuse as constituting proper training)[7] is invoked, although the last participant weighs the usefulness of the information (for safeguarding the general welfare) against the possibility of an unjust (but not extremely harmful) procedure.

Finally, it is important to notice the homogeneity of the responses to Slavery Generality. We attribute this result to the contemporary prototypicality of slavery as a transgression. The fact that participants reasoned uniformly in response to this prototypical scenario confirms the converse of our broad expectations that, in response to non-prototypical cases, participants would be guided by concerns not directly (and not necessarily) related to the moral domain.

5. Conclusion

Kelly et al. proposed that, in the context of "grown-up" scenarios, most adults would not categorize harm as moral wrongdoing. To the contrary, we found that most participants considered the harmful actions to be morally wrong, some considered them to be permitted, and still others may have even considered them to be morally right. Moreover, many participants raised concerns not directly (and not necessarily) related to the moral domain. We conclude by briefly addressing two potential problems with our argument and qualifying two of our claims.

The first potential problem has to do with the type of evidence utilized to test our hypothesis. Participants' justifications (and their coding) played a fundamental role in our characterization of what constitutes evidence, but recent research suggests that justifications are often an unreliable source of information on the cognitive processes underlying moral judgment (e.g., Cushman, Young, & Hauser, 2006; Haidt, 2001; Hauser, Cushman, Young, Jin, & Mikhail, 2007; Wheatley & Haidt, 2005). Thus, one might call into question the empirical ground supporting our hypothesis. Nonetheless, in responding to the specific scenarios of our research, participants' justifications did not appear to evince the common symptoms of unreliability. Participants did not show difficulties in articulating the reasons for their OK or Not-OK judgments, there were no apparent contradictions between justifications and patterns of judgments, and when participants gave multiple justifications, there was an internal coherence that does not suggest confabulation. Thus, we do not consider our evidence to be unreliable.

However, we do acknowledge that one should be cautious in accepting justifications as the main type of evidence for underlying psychological processes and that additional research should be pursued in order to give a better support to the claims raised here. Our data suggests many interesting possibilities to be explored in future research by directly manipulating features of the scenarios rather than relying primarily on justifications. For example, one might manipulate "informed consent" directly within the harmful scenarios. This approach would enable a more precise delineation of the extent to which specific components, such as informed consent, relate to questions of injustice and rights violations, and perhaps a more specific account of the relationship between harm and morality.

The second potential problem has to do with the status of our hypothesis in itself. Our hypothesis states that transgressions involving harm *and* injustice or rights violation are considered to be authority independent and general in scope. However, one may argue that, since authority independence and generalizability are part of the definition of rights violation, our hypothesis begs the question. If our objective were simply one of *a priori* conceptual analysis, we would agree. However, our objective was an empirical one about the components of the folk concept *moral transgression involving a person being subjected to harm*, and our hypothesis was about these components. Our results support the hypothesis that participants indeed reason about harm with sensitivity to the following components: *wrongdoing*, *harm*, *injustice*, *rights violation*, *authority independence*, *generalizability*. More generally, the findings of the Turiel tradition *generated by the moral/conventional task* suggest that human beings reason in this way very early in life and across societies.

The first qualification has to do with our claims about the components of the concept *moral transgression involving a person being subjected to harm*. We did not propose any explicit or detailed model of how these components relate to the structure of this concept. There may be other components involved. For example, a notion of *objective wrongdoing* may be an additional component (see Nichols (2004a) for an interesting discussion and counterevidence). Furthermore, we are not committed to a particular model of conceptual structure or to the position that all these components are part of the structure of this concept (for discussion of the different theories of conceptual structure in the cognitive sciences, see Laurence & Margolis, 1999; Rosch, in press). Our aim in this article was simply to elaborate a general position on the conceptual relation between harm and moral wrongdoing vis-à-vis alternative theoretical positions.

[6] In grouping punishment cases of informed consent with the other training cases under the same category, our coding scheme may be disregarding fundamental differences between the two cases. In this respect, it may benefit from revision.

[7] These are cases where participants adhered to their original informational assumptions about efficacy. There were similar cases when participants denied the utility of the actions in the context of No–No answers: "Negative reinforcement does not work" (Whipping Authority), "The suspect will be confused and therefore may answer questions inaccurately" (Prisoner Authority), "The physical pain although vanishing in a short period of time, destroys the ability of the soldiers to see options and think of their decisions" (Training Authority).

The second qualification has to do with our claim that harm (as pain or suffering) is not sufficient for morality. We do not claim that harm is irrelevant to morality in general. As Nichols (2004b) has persuasively argued, certain emotional reactions to the perception of harm may play an important role in the stabilization of norms proscribing harm, and over time may bias normative systems to increasingly categorize harmful actions as forbidden and, we would argue, as *morally* wrong in the sense we have discussed. Our claim is simply that more than the perception of a person being subjected to harm is required for the activation of the concept of *moral transgression involving a person being subjected to harm*.

Acknowledgments

We would like to thank three anonymous reviewers and the Editor for their helpful criticisms and suggestions, Daniel Kelly, Steven Stich, Kevin Haley, Serena Eng and Daniel Fessler for providing us with their raw data, the AHRC *Culture and the Mind* project for bringing to our attention some of the issues discussed in this article and Natalie Emmons for her assistance in this research.

References

Cushman, F., Young, L., & Hauser, M. D. (2006). The role of conscious reasoning and intuition in moral judgment: testing three principles of harm. *Psychological Science, 17*(2), 1082–1089.

Davidson, P., Turiel, E., & Black, A. (1983). The effect of stimulus familiarity on the use of criteria and justifications in children's social reasoning. *British Journal of Developmental Psychology, 1*, 49–65.

Fiddick, L. (2004). Domains of deontic reasoning: resolving the discrepancy between the cognitive and moral reasoning literatures. *The Quarterly Journal of Experimental Psychology, 57*, 447–474.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834.

Haidt, J., Koller, S., & Dias, M. (1993). Affect, culture and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology, 65*, 613–628.

Hauser, M. D., Cushman, F., Young, L., Jin, R. K.-X., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language, 22*(1), 1–21.

Helwig, C., Hildebrandt, C., & Turiel, E. (1995). Children's judgments about psychological harm in social context. *Child Development, 66*(6), 1680.

Kelly, D., Stich, S., Haley, K., Eng, S., & Fessler, D. (2007). Harm, affect, and the moral/conventional distinction. *Mind and Language, 22*, 117–131.

Laurence, S., & Margolis, E. (1999). Concepts and cognitive science. In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings* (pp. 3–81). Cambridge, MA: The MIT Press.

Nichols, S. (2004a). After objectivity: An empirical study of moral judgment. *Philosophical Psychology, 17*, 5–28.

Nichols, S. (2004b). *Sentimental rules: On the natural foundations of moral judgment*. New York: Oxford University Press.

Nucci, L. (2001). *Education in the moral domain*. Cambridge: Cambridge University Press.

Nucci, L., & Turiel, E. (1978). Social interactions and the development of social concepts in preschool children. *Child Development, 49*, 400–407.

Rosch, E. (in press). Concepts. In P. C. Hogan (Ed.). *The Cambridge encyclopedia of the language sciences*. New York: Cambridge University Press.

Smetana, J. (1981). Preschool children's conceptions of moral and social rules. *Child Development, 52*, 1333–1336.

Smetana, J. (1985). Preschool children's conceptions of transgressions: Effects of varying moral and conventional domain-related attributes. *Developmental Psychology, 21*, 18–29.

Smetana, J. (1986). Preschool children's conceptions of sex-role transgressions. *Child Development, 57*, 862–871.

Smetana, J. (1993). Understanding of social rules. In M. Bennet (Ed.), *The development of social cognition: The child as psychologist*. New York: Guilford Press.

Smetana, J., & Braeges, J. (1990). The development of toddlers' moral and conventional judgements. *Merrill-Palmer Quarterly, 36*, 329–346.

Smetana, J., Schlagman, N., & Adams, P. (1993). Preschool children's judgments about hypothetical and actual transgressions. *Child Development, 64*, 202–214.

Sousa, P. (2009). On testing the 'moral law'. *Mind & Language, 29*, 209–234.

Tisak, M. (1995). Domains of social reasoning and beyond. In R. Vasta (Ed.). *Annals of child development* (Vol. 11). London: Jessica Kingsley.

Tisak, M., & Turiel, E. (1984). Children's conceptions of moral and prudential rules. *Child Development, 55*, 1030–1039.

Turiel, E. (1983). *The development of social knowledge*. Cambridge: Cambridge University Press.

Turiel, E. (2002). *The culture of morality*. Cambridge: Cambridge University Press.

Turiel, E., & Smetana, J. (1984). Social knowledge and social action. The coordination of domains. In W. M. Kurtines & J. L. Gewirtz (Eds.), *Morality, moral behaviour, and moral development: Basic issues in theory and research* (pp. 261–282). New York: Wiley.

Wainryb, C. (1991). Understanding differences in moral judgments: The role of informational assumptions. *Child Development, 62*, 840–851.

Wainryb, C. (1993). The application of moral judgments to other cultures: Relativism and universality. *Child Development, 64*, 924–933.

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science, 16*(10), 780–784.

Weston, D., & Turiel, E. (1980). Act-rule relations: Children's concepts of social rules. *Developmental Psychology, 16*, 417–424.