Cognition 113 (2009) 93–97

Contents lists available at ScienceDirect

Cognition

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

Discussion

On the Morality of Harm: A response to Sousa, Holbrook and Piazza ☆

Stephen Stich [a,*], Daniel M.T. Fessler [b], Daniel Kelly [c]

[a] Rutgers University, Philosophy Department and Center for Cognitive Science, New Brunswick, NJ 08901-2882, United States
[b] UCLA, Anthropology Department and Center for Behavior, Evolution & Culture, Los Angeles, CA 90095-1553, United States
[c] Purdue University, Philosophy Department, West Lafayette, IN 47906-2098, United States

1. Introduction

The paper by Sousa, Holbrook and Piazza (SH&P) appears to have two distinct aims. First, it advances a number of criticisms of the argument developed in Kelly, Stich, Haley, Eng, and Fessler (2007), which was aimed at showing that a cluster of claims often attributed to Elliott Turiel and his followers are mistaken. Second, it proposes a new hypothesis about how people think about harmful actions and presents some valuable new data in support of that hypothesis. We will address each of these in turn.

We should begin by saying that we welcome the new data presented by SH&P. However, we do not agree that those data pose a problem for the argument advanced by Kelly et al. Indeed, we think the new data *support* the central claim made in Kelly et al. concerning how people think about transgressions in which someone is harmed. We suspect that SH&P may have misunderstood Kelly et al.'s argument, and that this misunderstanding, rather than any substantive disagreement, underlies many of their critical comments about the Kelly et al. paper. In Section 2, we will do our best to correct this misunderstanding by making clear exactly what Kelly et al. do (and do not) claim. We will then explain why we view the SH&P data as providing further support for Kelly et al.'s critique of the Turiel-inspired hypothesis they focused on. We will also briefly address SH&P's contention that Kelly et al. have misinterpreted the view of Turiel and his associates.

In Section 3, we will turn our attention to the new hypothesis proposed by SH&P. Our main theme, in that section, will be that SH&P have failed to take note of an important distinction between two kinds of rights, and that when the distinction is duly noted, their hypothesis is best viewed as a claim about the *rationality* of people's thinking about one class of harm transgressions. Once this point has been made, it becomes clear that there is an obvious way to generalize SH&P's hypothesis well beyond the domain of harmful transgressions. Ironically, if the generalized hypothesis turns out to be true, then harm is not playing a significant role in the sorts of judgments that are the focus of SH&P's study.

In Section 4, we will address SH&P's critique of the way in which Kelly et al. present their data. Though we agree that SH&P's reanalysis is helpful, we argue against their contention that Kelly et al.'s analysis fails to address the crucial feature of the data. We will end the section, and the paper, by reflecting on what might explain the differences between the data Kelly et al. report and the data presented by SH&P.

2. Kelly et al.'s critique of Turiel, and why SH&P's data supports that critique

Kelly et al. (2007) maintained that the majority of investigators in the Turiel tradition would endorse something like the following cluster of hypotheses:

(H1).

(i) "In moral/conventional task experiments subjects typically exhibit one of two *signature response patterns*. In the *signature moral pattern* rules are judged to be authority independent and general in scope. . .."[1] In the *signature conventional pattern* rules are judged to be authority dependent and not general in scope. . .. Moreover, these signature response patterns are what philosophers of science sometimes call

☆ We are grateful to two anonymous reviewers for their helpful comments on an earlier draft of this paper.

* Corresponding author. Tel.: +1 732 932 9091; fax: +1 732 932 8617.
E-mail address: sstich@rucs.rutgers.edu (S. Stich).

[1] The two other components of the "signature moral pattern" characterized by Kelly et al. are "violations are more serious, and rules are justified by appeal to harm, justice and rights". Following SH&P's lead, we will ignore these two in the current discussion.

0010-0277/$ - see front matter
doi:10.1016/j.cognition.2009.06.013

'nomological clusters' – there is a strong ('lawlike') tendency for the members of the cluster to occur together.

(ii) Transgressions involving harm, justice, or rights evoke the signature moral pattern.

(iii) Transgressions that do not involve harm, justice, or rights evoke the signature conventional pattern" (Kelly et al., 2007, 119–120; numbering modified).

We went on to argue that (H1) is mistaken. The argument for that conclusion had several parts. First, we cited a number of well know studies by Haidt, Koller, and Dias (1993), Nichols (2002), Nichols (2004), Nisan (1987) and others, showing that some transgressions that do not involve harm, justice or rights do not evoke the signature conventional pattern. These studies make it clear that (iii) is false. Moreover, in these studies transgressions not involving harm, justice or rights sometimes evoke *both* components of the signature moral pattern and sometimes evoke *only one* component. So the nomological cluster claim in (i), which maintains that there is a lawlike tendency for the components of a signature response pattern to be elicited together or not at all, is false as well. However, we could find nothing in the literature that was incompatible with (ii), hence we designed an investigation to test that claim.

One charge that SH&P make against Kelly et al. is that (ii) does not accurately capture the view of Turiel and his associates. However, for a number of reasons, we do not think that it would be productive to debate this issue in the present response. First, it is clear that the views advanced by Turiel and his associates have undergone some significant changes over the last 25 years. And that, of course, is entirely appropriate. In any flourishing scientific research tradition, one expects researchers to modify their views as new evidence is uncovered and new objections are proposed. Second, researchers in what SH&P refer to as "the Turiel tradition" do not speak with one voice. It would be rather worrisome if they did. Third, both Turiel and some of his followers have adopted a jargon that is often not clearly explained, and because of this it is sometimes hard to know exactly *what* they are claiming. Unfortunately, SH&P are of little help here since, while they insist that "Turiel and associates' position is more nuanced than outlined by Kelly et al." (SH&P, page 82), they offer no clear statement of what they take the Turiel hypothesis to be. Rather, they do something else; in their own words, they "*reinterpret* Turiel's hypothesis" (SH&P, page 84, our italics), and as they note themselves, this reinterpretation "departs from the way Turiel and associates frame the discussion" (SH&P, page 83). We think that SH&P's alternative hypothesis is an interesting one, and we will discuss it in some detail in Section 3. Our present point, however, is that SH&P's hypothesis is, by their own admission, an *alternative*, and so not relevant to the issue of whether or not (ii) accurately represents the view of Turiel and his associates. If SH&P wish to pursue this point, we submit that the discussion would best be served if they provided a clear and explicit account of what they take Turiel's nuanced hypothesis to be, along with the textual evidence supporting their formulation. When that has been done, we will be happy to revisit the question of whether or not Turiel and his associates are committed to (ii).

Though it is certainly possible to question whether (ii) accurately captures Turiel's view, it is clear that a number of influential psychologists and philosophers have interpreted Turiel and his associates as having established something like (ii). (Nichols (2002), Nichols (2004) are particularly clear examples.) Since the claim that transgressions involving harm evoke the signature moral response pattern has been very influential, it is important to know whether or not it is true. We now turn to this question.

According to SH&P, "the general hypothesis that harmful transgressions evoke the moral signature" is "vague" (SH&P, page 83). We disagree. What this hypothesis claims is that if a person takes an act to be a transgression[2] and if that person also recognizes that someone is harmed by that act, then she will judge the wrongness of the act to be authority independent and general in scope. It is hard to imagine a psychological hypothesis that is *less* vague! The reason SH&P think the hypothesis is vague seems to be that it "does not clarify what is supposed to establish that harmful actions are transgressions" (SH&P, page 83) – i.e., it does not say *why* the participant judges the action to be a transgression. This is quite correct. What the hypothesis claims is that *any* action that is judged to be a transgression (*for whatever reason*) and that is also judged to be harmful will evoke the moral signature. Thus one strategy we used to construct scenarios designed to test this hypothesis was to describe cases of harmful actions that participants might classify as transgressions for reasons that had little or nothing to do with the fact that they were harmful. If (H1) (ii) is true, cases such as this should evoke the signature moral pattern. But both our study and SH&P's partial replication found many subjects who judged a harmful action to be a transgression but did not exhibit the signature moral pattern. The "No–Yes" column in SH&P's Table 2 summarizes the relevant data. If (H1) (ii) were true, all the entries in that column would be close zero. As Table 2 makes clear, this was not the case either in our study or in the SH&P study. Since (H1) (ii) was the only component of (H1) that had not already been shown to be false, our study, taken together with the very useful data resulting from SH&P's partial replication, goes a long way toward establishing that every component of (H1) is false. This is exactly the conclusion that Kelly et al. defended.

SH&P's rather negative tone when discussing the Kelly at al. paper strongly suggests that they think they have refuted some view that Kelly et al. were defending, and, in the first sentence of their conclusion, they make clear what view they have in mind. "Kelly et al.," they maintain, "proposed that in the context of 'grown-up' scenarios, most adults would not categorize harm as moral wrongdoing." (SH&P, page 91) SH&P offer no textual evidence for this claim, and this is not surprising, since *we proposed no such thing*! What we did claim in our paper, clearly, explicitly and repeatedly, is that (H1) (ii) is false. Both our data and the data reported by SH&P support that claim.

[2] In this literature, the evidence that a person takes an act to be a transgression is that she says *no* to questions such as "Is the act OK?" or *yes* to questions such as "Is the act wrong?".

3. The SH&P hypothesis

Though SH&P criticize what they mistakenly take to be our hypothesis, most of their paper is devoted to defending a hypothesis of their own, which they characterize as "closely aligned with, but somewhat different from, the Turiel tradition" (SH&P, page 81). It is hard to judge how close their hypothesis is to the one they think the Turiel tradition defends, since, as we noted earlier, they offer no clear and explicit statement of what they take the Turiel hypothesis to be. Nevertheless, they do offer an explicit statement of their own hypothesis:

> [W]e reinterpret Turiel's hypothesis as follows: transgressions involving harm *and* injustice or rights violations 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 (SH&P, page 84).

It is important to see that when SH&P talk about harmful transgressions that are "perceived to entail injustice or rights violations," what they mean is that the harmful act is judged to be a transgression *because* it is an injustice or rights violation. This emerges very clearly when they set out the prediction they make based on their hypothesis:

> Whenever a participant answers Not-OK to the permissibility probe *and* their answer is *driven by* concerns with justice or rights... the answer to the moral signature (authority contingency or generality) probe will be Not-OK as well based on the same concerns (SH&P, page 84, second emphasis added).

We have two concerns with this hypothesis. To explain them, something needs to be said about the key terms 'rights' and 'justice'. We will focus on rights, though similar points can be made about justice. In both philosophy and jurisprudence, there is an enormous literature aimed at clarifying the notions of rights and justice (Nickel, 2006; Rawls, 1971; Waldron, 1984; Wenar, 2007). One crucial observation made in this literature is that some rights and some principles of justice are geographically and temporally local, while others may be universal. For example, in contemporary New Zealand, 15-year-olds have the right to have a driver's license, provided that they have passed the driving test. In India, by contrast, the minimum driving age is 18, and 15-year-olds have no right to have a driver's license, even if they could pass the test. Moreover, it is clear that these rights are dependent on the relevant authorities. The New Zealand parliament could raise the minimum driving age to 16, and if it did 15-year-olds would no longer have the right to have a driver's license.

In contrast with this, many moral and legal theorists, and many ordinary folks, think that *some* rights are universal and cannot be abridged by any authority. The *Universal Declaration of Human Rights*, adopted by the United Nations General Assembly in 1948, provides a long list of these "universal and inalienable" rights, including the right to own property (Article 17) and the right to education (Article 26). It is certainly not the case that *all* philosophers and legal theorists believe in the existence of universal and inalienable rights. Indeed, Jeremy Bentham famously derided the idea as "nonsense upon stilts" (Bentham, 1796). But whether or not such rights really exist, it is clear that many people, particularly people in Western democracies, *believe* that some rights are universal and inalienable. It is also clear that most people in Western democracies recognize that some rights, like the right to have a driver's license at age 15, are neither universal nor inalienable.

With this as background, let us return to SH&P's hypothesis and their prediction. If someone judges a harmful act to be a transgression, SH&P tell us, and if they make that judgment because they think the harmful act is an injustice or a rights violation, then they will judge that the transgression is authority independent and general in scope. But what we have just seen is that many people in Western democracies think there are two kinds of rights – those that are universal and inalienable, and those that are not. Which sort of rights do SH&P have in mind? On what we take to be the most charitable interpretation of their view, the answer is the universal and inalienable kind. But if that is what they intend, then it appears that what their hypothesis is really claiming is that when people judge a harmful action to be a transgression because they believe it violates a universal and inalienable right, their answers to questions about authority independence and generality are *rational*.

To see the point, let us imagine an experimental participant, Tom, who believes that the act described in an experimental scenario is harmful, and judges that it is a transgression. (He says it is Not-OK.) Moreover, suppose that Tom judges the act to be a transgression *because* he believes it to be a violation of a universal and inalienable right. We now ask him whether the act would be OK if some authority says it is OK. Clearly, unless Tom is seriously irrational, he will say no, since inalienable rights cannot be altered or suspended by any authority – that, after all, is what it *means* for a right to be *inalienable*. Next we ask him whether the act would be a transgression in some distant land or at some other time in history. Once again, unless he is seriously irrational, he will say no, since what it *means* for a right to be *universal* is for it to obtain at all times and in all places. Thus, on what we take to be the most charitable interpretation of SH&P's hypothesis, it appears that its only real empirical substance is the claim that in tasks like the ones we have described, most participants are not seriously irrational – they give the answers that follow, by simple logic, from what they believe.

Aristotle famously maintained that man is a rational animal. So neither Aristotle nor those who agree with him would find this hypothesis to be particularly surprising. However, in recent decades, the literature on reasoning, judgment and decision-making has made it abundantly clear that Aristotle was far too optimistic. In many reasoning, judgment and decision-making tasks many people are *not* rational (Gilovich, Griffin, & Kahneman, 2002; Kahneman, Slovic, & Tversky, 1982;

Samuels & Stich, 2004). So, if the SH&P hypothesis is true, it is certainly of some interest, since it points to an important domain in which people buck the trend revealed in much recent work – a domain in which people actually do reason correctly.

But now let's look more carefully at what that domain is, and consider what role, if any, *harm* has to play in characterizing it. At first, harm seems to be fairly central to the hypothesis, an impression that is certainly supported by SH&P's presentation of their work. We think this is misleading, however. If we understand them correctly, SH&P's hypothesis claims that:

(1) If a person judges a harmful act to be a transgression, and if they make that judgment because they think the harmful act is a violation of a universal and inalienable right, then (in accordance with basic principles of rationality) they will judge that the transgression is authority independent and general in scope.

Stated this way, an obvious generalization of the hypothesis suggests itself. To generate the generalization we need only drop the restriction to *harmful* acts, producing the following:

(2) If a person judges *any* act to be a transgression, and if they make that judgment because they think the act is a violation of a universal and inalienable right, then (in accordance with basic principles of rationality) they will judge that the transgression is authority independent and general in scope.

We are inclined to think that if (1) is true, it is because (1) is a special case of (2). But, for the moment, this is simply a speculation. Further research will be needed to determine whether this speculation is true. If it is, then, ironically, SH&P's paper, "The Morality of Harm", will have told us nothing distinctive about moral judgments that involve harmful acts. Rather, the take-home message will be that people are not egregiously irrational in reasoning about the authority independence and universal applicability of acts which they believe to be violations of universal and inalienable rights – *whether or not those acts are harmful*!

4. SH&P's results and methods

One of SH&P's principal criticisms of Kelly et al. is that their analysis and presentation of results fail to address a crucial feature of the data. Kelly et al. explored the possibility that a nontrivial fraction of participants who judged a behavior, described as near in space and time, to be unacceptable (answering "No" to the "Is it OK?" question) would find the same behavior acceptable (answering "Yes" to this question) when it was described as either approved by an authority or distant in space or time. For each dyad of scenarios, there were thus four possible pairs of responses, namely (a) [No + No]; (b) [No + Yes]; (c) [Yes + No]; and (d) [Yes + Yes]. SH&P sharply criticize Kelly et al.'s analysis, arguing that, because Kelly et al. pooled responses across participants, it is impossible to determine whether or not a substantial portion of participants responded [No + Yes], as Kelly et al. claim. Instead, SH&P revisit Kelly et al.'s results (see SH&P Table 2), using the above four categories to organize the data on the basis of individual participants. While it is interesting to view the results in this manner, SH&P are simply mistaken in arguing that Kelly et al.'s use of pooled results made it impossible to conclude that a nontrivial proportion of participants who disapproved of behavior in one context also approved of it in another. To determine whether a nontrivial proportion of participants follow the (b) pattern, all that is needed is to compare the percentages for each question. Consider, for example, the results for the Whipping Generality scenarios: in the pooled sample, approximately 90% of participants answer "No" to the question of the first type, and 50% of participants answer "Yes" to the question of the second type. Since 90% answered "No" to the question of the first type, at most 10% of participants could possibly have responded in pattern (d). Thus any figure above 10% for "Yes" to the question of the second type can only be attributed to pattern (b), i.e., even a highly conservative interpretation places this figure at 40%. Hence, while Kelly et al.'s presentation of results may have demanded more of the reader than does SH&P's presentation, SH&P are wholly incorrect in claiming that Kelly et al.'s analysis could not answer the question that their enterprise was intended to address.

Of course, the fact that there is no logical need to present results in a participant-specific manner in order to examine the pattern at issue does not mean that such a presentation is not of potential interest. We turn, therefore, to SH&P's Table 2. Summarizing this table, SH&P state "It is important to notice that ... except for Whipping Generality, the results of the replication are quite similar to Kelly et al.'s results." Inspection of the crucial No–Yes column of this table makes it difficult to see how SH&P arrive at this conclusion. Granted, SH&P's percentages are identical to Kelly et al.'s with regard to the Training scenario, and nearly identical with regard to the Prisoner scenario. However, in addition to the Whipping Generality scenario (for which 16% of SH&P's participants provided the No–Yes pattern, in contrast to 41% of Kelly et al.'s participants), they also diverge somewhat in the Slavery scenario (0% versus Kelly et al.'s 5%), and diverge substantially in the Whipping Authority scenario (3.0% versus Kelly et al.'s 17.5%). Hence, there are notable differences between the two studies in regard to results for three of the five pairs of scenarios. What are we to make of such differences? SH&P recruited participants and administered their instrument in ways nearly identical to those employed by Kelly et al., suggesting that, all else being equal, one might reasonably expect comparable results. However, all else is not equal – as evident in Table 2, SH&P's samples are between 1/5 and 1/6 the size of those of Kelly et al. for each item. So one possible explanation for the differences between the

two studies might well be that SH&P's samples were much smaller. Further research will be needed to determine whether that is the correct explanation.

References

Bentham, J. (1796). Anarchical Fallacies. In Waldron (1987), pp. 46–76.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge: Cambridge University Press.

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.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

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

Nichols, S. (2002). Norms with feeling: Toward a psychological account of moral judgment. *Cognition, 84*, 223–236.

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

Nickel, J. (2006). Human rights. In Edward N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2006 Edition), URL: <http://plato.stanford.edu/entries/rights-human/>.

Nisan, M. (1987). Moral norms and social conventions: A cross-cultural comparison. *Developmental Psychology, 23*, 719–725.

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.

Samuels, R., & Stich, S. (2004). Rationality and psychology. In Alfred Mele & Piers Rawling (Eds.), *The Oxford handbook of rationality* (pp. 279–300). Oxford: Oxford University Press.

Waldron, J. (Ed.). (1984). *Theories of rights*. Oxford: Oxford University Press.

Waldron, J. (Ed.). (1987). *Nonsense upon stilts: Bentham, Burke, and Marx on the rights of man*. London: Methuen.

Wenar, L. (2007). Rights. In Edward N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2007 Edition), URL: <http://plato.stanford.edu/entries/rights/>.