Running Head: UTILITARIAN HARM

Authority Dependence and Judgments of Utilitarian Harm

Jared Piazza*

University of Pennsylvania, Department of Psychology

Paulo Sousa

Queen's University Belfast, Institute of Cognition and Culture

and

Colin Holbrook

University of California, Los Angeles, Department of Anthropology

April 30, 2013

Word Count: 8,069

*__Corresponding author__: Jared Piazza, Department of Psychology, University of Pennsylvania

3720 Walnut Street, Solomon Labs Bldg., Philadelphia, PA 19104 USA Tel: 215-898-7866

Email: jpiazza@psych.upenn.edu

Abstract

Three studies tested the conditions under which people judge utilitarian harm to be *authority dependent* (i.e., whether its right or wrongness depends on the ruling of an authority). In Study 1, participants judged the right or wrongness of physical abuse when used as an interrogation method anticipated to yield useful information for preventing future terrorist attacks. The ruling of the military authority towards the harm was manipulated (prohibited vs. prescribed) and found to significantly influence judgments of the right or wrongness of inflicting harm. Study 2 established a boundary condition with regards to the influence of authority, which was eliminated when the utility of the harm was definitely obtained rather than forecasted. Finally, Study 3 replicated the findings of Studies 1-2 in a completely different context—an expert committee's ruling about the harming of chimpanzees for biomedical research. These results are discussed as they inform ongoing debates regarding the role of authority in moderating judgments of complex and simple harm.

Keywords: Utilitarian harm; authority; moral judgments; moral reasoning; moral dilemmas; moral/conventional task

Authority Dependence and Judgments of Utilitarian Harm

While most people agree that it is wrong to intentionally cause another person pain or suffering, people also recognize that there are circumstances in which harming someone may be justified. Though there may be disagreement about what qualifies as an adequate justification for harm (Gert, 2004), in general, people seem to relax their condemnation when harmful acts are performed with the intention of producing *utility*, that is, a greater good, such as the alleviating of even greater suffering (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Nichols & Mallon, 2006). But how exactly do people balance utility and the causation of pain or suffering in their judgments of utilitarian harm? Could the sanction or proscription of a recognized authority make a difference in these judgments?

Numerous psychological studies conducted by Turiel and his colleagues have shown that adults and children consistently condemn acts that cause pain or suffering, and reject the notion that any authority figure can undo the impermissibility of such harm (Davidson, Turiel, & Black, 1983; Laupa & Turiel, 1986, 1993; Nucci, 2001; Nucci & Turiel, 1978, 1993; Smetana, 1981, 1985, 1993; Tisak & Turiel, 1984; Turiel, 1983; Weston & Turiel, 1980). However, the focus of this research has been on cases of harmful actions that clearly involve injustice and rights violations, where the causation of pain or suffering is seen as motivated exclusively by selfish reasons—for example, an innocent child is pushed off a swing or is hit by another child just for fun. Such cases exemplify what we call *simple harm* (others have called these cases "prototypical" violations; e.g., Wainryb, 1991). Rarely have psychologists from this cognitive-developmental tradition investigated the way people reason about cases of *complex harm*, where the causation of pain or suffering is placed in conflict with other considerations, such as whether utility may be derived from the act, or whether the actor has other justifiable reasons for causing harm (however, see Turiel, Hildebrandt, & Wainryb, 1991; Wainryb, 1991, 1993, for notable

exceptions). Thus, the possibility remains that the policies of relevant authorities, which do not sway judgments of simple harm, do inform evaluations of complex harm, particularly when the possibility of utility is in question.

In contrast to this developmental tradition, though consistent with an even earlier tradition pioneered by Kohlberg (1969), there is a growing interest among moral psychologists, neuroscientists, and experimental philosophers in the psychological processes involved in reasoning about cases of complex harm, where the causing of pain or suffering does not occur solely for selfish reasons (e.g., Cushman, Young, & Hauser, 2006; Greene et al., 2001; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Koenigs, Young, Adolphs, Tranel, et al., 2007; Nichols & Mallon, 2006; Valdesolo & DeSteno, 2006). One common case used by researchers in this tradition is Foot's (1967) *trolley dilemma*. In this well-known scenario, the rule "one should not harm an innocent person" is placed in conflict with the pursuit of a greater good (saving a number of innocent lives). In this case, most people find it permissible for a person to kill an innocent in order to save the lives of five others (Cushman et al., 2006; Greene et al., 2001; Thomson, 1985), adopting a good-maximizing (or aggregate cost-benefit) solution to the dilemma. Although there are versions of this dilemma in which most people consider it wrong to adopt a good-maximizing solution to the dilemma (e.g., the *footbridge dilemma*, where an innocent must be physically pushed off a footbridge to stop a runaway trolley; Greene et al., 2001), it has been shown that when the consequences of not adopting such a solution are catastrophic (not simply the death of five innocents but of thousands of people), most people find it permissible to kill an innocent person to obtain a greater good (see Nichols & Mallon, 2006).

For the most part, the cognitive-developmental tradition pioneered by Turiel and his colleagues and the moral dilemma tradition have pursued separate trajectories. Whereas the

former probes whether the impermissibility of simple harm is considered to be independent of the permission of an authority, the latter probes whether complex harm is considered to be permissible, without concerning itself with the potential influence of an authority in modifying the perceived normative status of the harm. Recently, however, a few researchers have sought to integrate these traditions by asking whether people conceptualize the wrongness of complex harm as unchangeable by the decree of an authority or other contextual factors (Kelly, Stich, Haley, Eng, & Fessler, 2007; Sousa, 2009; Sousa, Holbrook, & Piazza, 2009; Stich, Fessler, & Kelly, 2009). Research into this question, however, has been hampered by theoretical disagreement and methodological limitations (see Sousa et al., 2009; Stich et al., 2009). First, it was not clear whether participants who changed their judgment according to the dictates of an authority did so out of concern for the authority's ruling in and of itself, or for orthogonal reasons, such as whether the authority possessed or lacked adequate knowledge about the probable utility of the harm (for details, see Sousa, 2009). Second, the variable of utility was not manipulated experimentally in these studies, and there was a great deal of variability in participants understanding of whether the harmful action was likely to produce utility or not (Sousa et al., 2009).

In this paper, we present new evidence from three experiments in which we manipulated the stance of an authority towards a particular class of complex harm—utilitarian harm—while assessing judgments of the harm. In addition to manipulating the ruling of an authority towards the harm, we also probed participants' understanding of the role authority played in their judgments. We show that, unlike cases of simple harm, where the normative status of the act is understood to be unalterable by an authority, many people do not understand utilitarian harm to be completely independent of an authority's influence. Rather, under *prospective* conditions of

anticipated utilitarian benefits, judgments of harm may be altered by the ruling of a legitimate authority.

The Present Hypotheses and Studies

We surmise that, for many people, utilitarian harm situations represent a *genuine moral conflict*—that is, respondents may be truly divided in their reasoning about the harmful act. On the one hand, they may recognize that the victim's rights would be violated by the harm, while on the other hand they may recognize that there is potential utilitarian value to the harm. For such conflicted individuals, for whom the rationales for and against committing the harmful act carry equal weight, the ruling of an authority may help tip the balance toward greater disapproval of the act when prohibited, or greater approval when prescribed.

To test this hypothesis, we conducted three experiments examining the role of authority in judgments of utilitarian harm. In each study, we adopted a between-subjects experimental methodology where we manipulated the ruling of a legitimate authority towards an act of utilitarian harm. In Studies 1-2, participants were presented an adapted version of the military interrogation scenario from Kelly et al. (2007) and Sousa et al. (2009), in which a military officer performs an act of harm (in the present case, an act of physical harm) in the pursuit of utilitarian benefits (to obtain information from a terrorist suspect that could save lives). Across Studies 1-2, the stance of a legitimate authority (military law[1]) was manipulated, such that the authority either prescribes or prohibits the use of the harmful interrogation methods, while we held constant perceptions of the utility of the harm. In Study 3, we extended the investigation to a completely different authority context—an Institutional Animal Care and Use Committee that

[1] Literally speaking, "military law" is not an authority figure or social body. However, we use military law as a reasonable proxy of an authority figure.

either "approves" or "rejects" a scientist's proposal to damage the brains of healthy chimpanzee subjects as a necessary component of an experimental biomedical procedure that could produce treatments for neurological disorders.

Across all three studies, we also probed participants' perceptions of the influence of authority on their judgments to determine whether these perceptions reflect a concern for *authority normativity*—i.e., the perceived right or wrongness of the act reflects a concern for the authority's ruling being adhered to or violated—rather than other influences authority might entail (see Sousa et al., 2009). In Studies 1-2, we investigated this possibility by directly probing participants about the role of authority in their decisions. In Study 3, we assessed authority influence less directly through open-ended responses.

Finally, in Studies 2-3, we sought to establish a boundary condition on the influence of authority on judgments of utilitarian harm. We hypothesized that the influence of authority would be restricted to cases where the harmful act was anticipated to produce utility, but the utility itself had yet to be realized (in contraposition to cases where the utilitarian outcome has already been obtained). In cases of unrealized utility, respondents can only be hopeful that the harm will produce the projected benefits (e.g., saving lives or alleviating greater suffering), but cannot be certain. Without definitive proof that the beneficial outcome will be realized, the possibility remains that the harm will occur unnecessarily or in vain. We reasoned that conditions of expected utility without definitive proof are the ideal conditions for the normative force of authority to exert an influence on respondents' judgments, since it is under these prospective conditions that a respondent is likely to experience the greatest ambivalence concerning the justifiability of the harmful act. By contrast, in cases where there is clear knowledge that the utilitarian outcome has been obtained, we do not expect respondents to be

swayed as much by the ruling of an authority, since these definite conditions should enhance the conviction that the harmful act was justified.

Study 1

The main aim of Study 1 was to establish the causal role of authority in judgments of utilitarian harm by manipulating the ruling of an authority within a between-subjects design. In earlier studies by Kelly et al. (2007) and Sousa et al. (2009), judgments of the harmful act were assessed using a dichotomous choice ("OK" or "Not-OK") modelled after the moral/conventional task (i.e., the within-subjects interview methodology used by Turiel and colleagues to discriminate moral and conventional transgressions, and to probe for authority contingency; e.g., Tisak & Turiel, 1984). Consistent with our proposition that individuals may be divided in their position towards acts of utilitarian harm, due to *an equal weighting of rationales*, we included a third option assessing participants who were "truly divided" in their reasoning regarding the harmful act, in addition to providing them with "right" or "wrong" options. We also moved beyond these earlier studies by probing whether or not participants' perceived the ruling of an authority *in and of itself* as having an influence on their judgment. Given research suggesting that individuals are not always conscious of the input that affects their judgments (e.g., see Cushman et al., 2006; Hauser, Cushman, Young, Jin, & Mikhail, 2007; Wheatley & Haidt, 2005), we directly probed participants' perceptions of the role authority played in their judgments.

Method

Participants. Participants were 163 adults (101 male, 62 female; M_{age} = 32.85 years, SD = 12.21), who were recruited via Amazon's Mechanical Turk (www.mturk.com; see Buhrmester,

Kwang, & Gosling, 2011) in exchange for $.50 payment. The nationality of the sample was 59% American, 36% Indian, and 5% other or multiple nationalities.

Materials and procedures. Participants read one of two versions of the vignette. They received instructions to read all the information carefully, to think about the vignette as if it were an actual event, and to treat the information presented as if it were factually accurate. These instructions were meant to dissuade participants from challenging the informational assumptions of the scenario (e.g., that torture is a reliable means of procuring accurate information—an issue that is hotly debated; see Brumfiel, 2007; Janoff-Bulman, 2007). The vignette read:

> "Sergeant Johnson is interrogating a member of a terrorist organization who may possess information about future terrorist attacks. As a means of obtaining information from the terrorist about the attack, Sergeant Johnson uses threats and physical abuse during the interrogation. Although no one has ever been killed or permanently disabled by the physical abuse they received during these interrogations, suspects often end up with bruises or injuries that last for a week or more. Research has definitely shown that information obtained by such harsh interrogation methods is reliable, and [however] military law currently prescribes [prohibits] the use of physical abuse in interrogation."

After reading the vignette, participants responded to three questions. First, participants provided their personal judgment regarding Sergeant Johnson's use of harsh interrogation methods. They selected from the following three options, "Which option best reflects your personal opinion on the matter?": (1) *It is right to use such harsh interrogation methods to gain information from the terrorist in this situation*; (2) *It is wrong to use such harsh interrogation methods to gain information from the terrorist in this situation*; (3) *I'm truly divided. There are*

equally valid reasons for using and not using such harsh interrogation methods to gain information from the terrorist in this situation. Next, participants were probed regarding the influence of the authority on their evaluation: "Did the fact that Sergeant Johnson followed [violated] military law with his use of harsh interrogation methods influence your previous choice?" (*Yes, No, I'm not sure*). This measure was used to assess the extent to which a participant perceived their judgment of the harm to be affected by the ruling of the authority—the fact that it was followed or violated—in and of itself. Lastly, a third, partly free-response measure assessed whether participants perceived there to be any additional influences of authority on their judgments, independent of the adherence to or violation of authority rule: "Did the position of military law regarding the use of harsh interrogation methods influence your choice in any other manner (i.e., other than the manner described in the previous question)?" (*Yes, No, I'm not sure*). If they selected "Yes," to this question, they were asked to explain their response and were provided a textbox in which to respond. All participants were debriefed at the end.

Results and Discussion

Judgments of the harmful act. Table 1 presents the frequency of each judgment category as a function of authority condition. Though "right" was the modal response for both authority conditions, a Chi-square test revealed that the overall distribution of responses within the authority conditions was significantly dissimilar than what would be expected by chance alone, $\chi^2(2) = 6.89$, $p < .03$, $\phi_c = .21$.[2] Specifically, when the authority prescribed the use of

[2] The inclusion of non-American participants in the experiment was never intended by the researchers. However, since the effect of authority was even stronger with non-American participants excluded from the analysis, $\chi^2(2) = 11.72$, $p < .01$, $\phi_c = .34$, and thus the results

harmful interrogation methods, most participants judged the sergeant's actions to be right (55%), 37% were truly divided in their judgment, and a mere 8% thought the sergeant's actions were wrong. By contrast, when the authority prohibited the use of the harm, a larger percentage of participants (23%) deemed the sergeant's actions wrong, a lower percentage of participants viewed the sergeant's actions as right (48%), and a lower percentage (28%) were truly divided in their judgment.

[Insert Table 1 about here]

The perceived influence of authority rule. Overall, a slight majority of participants (56%) reported that the position of military law towards the use of harmful interrogations did not influence their judgment of the harm; 10% were unsure whether authority had such an influence, and about one-third of respondents (34%) reported a normative influence of authority on their judgment. The distribution of responses was equivalent across the two authority conditions, $\chi^2(2) < 1$, *ns*, mirroring the overall pattern of responses. Thus, the perceived normative influence of authority on judgments was equivalent across the two authority conditions.

Among those 34% who reported a normative influence of authority on their judgment, there was a significant difference in judgments as a function of authority condition, $\chi^2(2) = 12.73$, $p < .002$. In the authority prescribes harm condition, most participants who perceived an influence of authority judged the sergeant's actions to be right (68%), or were truly divided (25%), though a few (7%) judged the sergeant's actions to be wrong. In contrast, in the authority prohibits harm condition, slightly less than half of participants who perceived an influence of authority on their decision judged the harm to be wrong (44%), while 30% were truly divided,

---

differed only by a small degree with non-American participants included, the reported analyses involve the entire sample.

and 26% judged the sergeant's actions as right. In sum, participants who reported an authority influence largely provided authority-consistent judgments, though some were truly divided in their judgment as a result of the authority's position, and a select few provided authority-opposing judgments, though this occurred more when the authority prohibited the harmful act than when the authority permitted it (presumably because these participants tended to take issue with the authority's rule prohibiting harm that would likely save lives).

Finally, very few participants reported that the authority had an influence on their judgment beyond the sergeant following or violating military law (2-3%). Most of the responses participants offered to explain this selection either referenced the utility of the harm (e.g., "With simple questioning [the suspect] will not disclose any information") or injustice (e.g., "Military law can be used unfortunately for wrong reasons and innocent people get punished which puts human rights into peril"), though one respondent questioned whether military law was consistent with international law ("It is generally globally illegal to use torture methods"), which is arguably a concern about authority legitimacy. Overall then it is safe to conclude that when participants acknowledged a concern for authority, this concern focused almost exclusively upon the normative ruling of the authority and whether or not it had been followed, as opposed to other concerns about authority.

In Study 1, we demonstrated the effect of authority with regards to an act of physical harm with expected utilitarian benefits. Participants were less disapproving of the harm when military law prescribed the act than when military law prohibited it. Though many participants rejected the notion that the position of the authority influenced their judgment of the utilitarian harm, about one-third of participants acknowledged a concern for authority rule. Furthermore,

when participants reported a concern for authority, they almost exclusively reported a concern that the authority's ruling had been followed or violated.

Study 2

In Study 1, the utility derived from the harm was *prospective* (i.e., expected, but not yet attained). In Study 2, we sought to uncover a bound to the influence of authority on judgments of utilitarian harm. Here we tested the hypothesis that the influence of authority on judgments of utilitarian harm are limited to cases where utility may be expected, but has not yet been realized. We reasoned that authority should have little influence upon judgments of harm that have *already* yielded utility, as participants should generally agree that the harmful act was the right course of action, or at least be divided in their judgment, since the net benefits have definitely been obtained. Such definite attainment should eliminate any lingering doubts respondents may have about the projected utility of the harmful interrogations, and thus may provide stronger justification for the causation of harm, though even under such definite conditions we would not expect all concerns about the cruelty or perceived injustice of the harm to be eliminated. Thus, we predicted that only a small minority of individuals would judge harm that has clearly achieved utilitarian benefits to be outright "wrong", and this would be true irrespective of the authority's ruling, though we expected a sizable number of individuals to remain divided in their judgment of the act, given the harmful nature of the act.

Method

Participants. Participants were 134 adults (85 male; M_{age} = 31.32 years, SD = 12.08) residing in the U.S., who participated via the same Web service used in Study 1 (www.mturk.com) in exchange for $.50 payment; 85% were White/Caucasian, 10% Asian, and 5% other ethnicities.

Materials and procedures. The materials and procedures were similar to Study 1. However, the wording of the vignette was changed slightly to describe utility that was definitely obtained from the use of physical abuse in the interrogation. Participants read that, as a result of the harsh interrogation methods, the suspect revealed information that was essential for stopping the terrorist attack, and many innocent lives were saved as a result.[3] Again, the harmful act occurred when military law either prescribed (n = 66) or prohibited (n = 68) the use of physical abuse in prisoner interrogation. Participants responded to two measures: the same judgment probe and authority influence probe as before. However, this time the three options (*right, wrong, truly divided*) were worded in the past tense.

Results and Discussion

As predicted, judgments of the utilitarian harm were similar across the authority conditions, $\chi^2(2) = 1.44$, $p = .49$, $\phi_c = .10$, when the harm had definite utility. As can be seen in Table 2, in both conditions, the vast majority of participants endorsed the harm (42% overall) or

[3] By presenting participants with evidence that numerous lives were saved, we may have also incidentally increased the perceived scale of the utility, since the scale of the utility was not explicitly stated in Study 1. To test this possibility, in a separate vignette study (N = 148), we maintained the prospective nature of the utility while stating explicitly that the information gained from the terrorist via harmful interrogations could prevent a future terrorist attack "which could seriously threaten American lives". This study replicated the effect of authority on judgments of the harm, $\chi^2(2) = 7.15$, $p < .03$, $\phi_c = .22$, suggesting that increasing the magnitude of the utility (while maintaining the prospective nature of the utility) does not eliminate the effect of authority when the utility is expected, but not yet obtained. Of course, this does not rule out the possibility that if the prospect of not harming an individual were catastrophic (e.g., thousands or millions of innocents would die), people would resoundingly approve of the harm irrespective of authority. Thus, we accept that it is still possible that the scale of utility may undermine an authority influence in extreme situations.

were truly divided in their judgment (44% overall). Furthermore, equally few participants deemed the harm wrong within the two authority conditions (prescribe: 11%; prohibit: 18%).

Just less than half of the participants (44%) reported that the position of military law influenced their judgment of the harmful act, 9% were unsure, and 47% reported that the military law's stance had nothing to do with their judgment. These percentages were nearly equivalent across authority conditions, $\chi^2(2) < 1$, *ns.* Looking only at participants who said the authority influenced their judgment, the difference in response pattern was significantly dissimilar across authority conditions, $\chi^2(2) = 9.37$, $p < .01$. In the prescribe condition, among those who reported being influenced by the authority, 43% deemed the sergeant's actions right (an authority-consistent judgment), 54% were divided, and 3% said it was wrong. In the prohibit condition, 14% who reported an authoritarian influence judged the act right, 62% were divided, and 24% said the sergeant's actions were wrong (an authority-consistent judgment).

[Insert Table 2 about here]

In summary, when the utility of the harm was definitely obtained, the vast majority of participants deemed the harm right, or were divided in their judgments, across conditions of authority. Very few participants deemed the harm completely wrong, even in the authority prohibits condition. Finally, a large percentage of participants from the authority prohibits condition reported they were truly divided in their judgment as a result of the authority's stance towards the harm. This seems to suggest that concern for the rule of authority did not completely evaporate when the harm had definite utility, but, unlike in Study 1's scenario of prospective utility, the violation of authority was no longer reason to judge the act to be outright wrong.

Study 3

The goal of Study 3 was to demonstrate that our findings regarding authority dependence are not limited to a single harm context, target of harm, or class of authority. To this end, we recruited a new sample of participants to judge an act of utilitarian harm set within a context of biomedical research. The vignette described an act of physical harm inflicted on twenty healthy chimpanzees by a neuroscientist pursuing biomedical treatments for neurodegenerative illnesses such as Alzheimer's. The authority in the scenario was the Institutional Animal Care and Use Committee (IACUC) at the scientist's university, a body comprised of five animal research experts. Participants read that the scientist conducted the experiment with or without the IACUC's approval—thus, the authority's ruling was either followed or violated. We also manipulated the *definiteness* of the utility gained from the experiment, to test our full hypothesis within a 2x2 factorial design.

Study 3 deviated from Studies 1-2 by recruiting non-human subjects as the target of harm. Nevertheless, we surmised that chimpanzees would provide a suitable proxy for human subjects given that many people bestow to chimpanzees moral consideration at a level comparable to that of human lives (consistent with The Great Ape Project's declaration that great apes are our "community of equals" with certain inalienable rights; Singer & Cavalieri, 1993), due to their complex social and cognitive abilities, many of which they share in common with humans. Furthermore, since humans are protected from invasive biomedical testing, we thought chimpanzees would provide a more suitable target of harm in this context (note that at the time this study was conducted it was still legal for chimpanzees to be used in invasive biomedical research in the United States, though recent legislation aims to phase out invasive experimentation on chimpanzees over the next three years; Wadman, 2012). Still, there is much disagreement about the precise rights we should extend to chimpanzees; thus, we included in

Study 3 a measure to assess whether or not chimpanzees are perceived to have the right to not be harmed even for human benefit. Consistent with the notion that the perception of rights violations is essential to the condemnation of harmful acts (see Sousa et al., 2009; Sousa & Piazza, 2013), we predicted that participants who attributed to chimpanzees the right not to be harmed for human benefit would exhibit greater disapproval of the scientist's actions than those who do not extend such rights to chimpanzees.

Also departing from Studies 1-2, in Study 3 we used open-ended (free-response) justifications as a less direct method of examining participants' perception of the influence authority exerted on their judgments. Arguably, directly probing participants about the role of authority could lead some individuals to confabulate the influence of authority on their judgment after having their attention drawn to the fact that they made an authority-consistent or authority-inconsistent judgment (though the chances of confabulation are attenuated by the fact that the authority's ruling is central to the vignettes, thus drawing attention to this factor prior to the right/wrongness probe). The use of open-ended responses avoids this limitation, but carries its own limitations as well—namely, that participants may simply fail to report all the relevant input contributing to their judgment. Lastly, we included a check on the perceived legitimacy of the IACUC's authority ("how important is it that the authority's ruling be followed?"), to confirm that the authority was respected at equivalent levels across the experimental conditions.

Our primary hypothesis predicted that participants would condemn the harmful act more when the scientist violated the IACUC's ruling than when he followed it, but we expected this difference to emerge mainly when the utility was prospective, having yet to be obtained.

Method

Participants. Participants were 304 adults (201 male, 103 female; M_{age} = 30.25 years, SD = 10.65) residing in the United States, recruited through the same web service as in prior studies (www.mturk.com) in exchange for $.50 payment. Previous participants were excluded from participation. All 304 participants provided sensible responses to the scenario, and therefore were retained; the sample was 80% White/Caucasian, 20% other ethnicities.

Design. The design was 2 (*authority ruling*: rejects vs. approves) x 2 (*utility definiteness*: definite vs. indefinite) between-subjects factorial. Participants were randomly assigned to one of four conditions: rejects/definite (n = 75), rejects/indefinite (n = 76), approves/definite (n = 76), or approves/indefinite (n = 77).

Materials and procedures. All participants were instructed to think about the scenario as if the events described had actually occurred. For all participants, the vignette began:

> "Professor Anderson is a psychobiologist working on the frontiers of neuroscience in the U.S. His research is attempting to show that neural tissue can be removed from the healthy brains of chimpanzee fetuses and implanted into the brains of individuals suffering from degenerative brain disorders, such as Parkinson's and Alzeimer's, in order to treat their illness. However, in order to have test subjects for his experiments, he must first damage the healthy brains of living chimpanzees, by opening their skulls and making lesions (surgical cuts) to the cerebral cortex of their brains. This procedure causes permanent behavioral and memory impairments to these chimpanzees. Professor Anderson applied to the Institutional Animal Care and Use Committee (IACUC) at his university to get permission to conduct an experiment where he would damage the brains of twenty healthy chimpanzees. The IACUC is comprised of five animal research

> experts. It is responsible for reviewing and either approving or rejecting all animal research conducted at the university."

This was followed by a second part, in which the authority variable was manipulated:

> "The IACUC approved Professor Anderson's proposed experiment, because it has the potential to produce results that would be useful for developing treatments for human degenerative diseases and alleviating the suffering of many human beings. Thus, consistent with the IACUC's ruling, Professor Anderson carried out his experiment. [The IACUC rejected Professor Anderson's proposed experiment because of the harm it would cause to twenty healthy chimpanzees. Nevertheless, against the IACUC's ruling, Professor Anderson carried out his experiment.]"

Finally, participants assigned to the definite utility condition read the additional sentence:

> "As a result, Professor Anderson had a successful breakthrough in developing treatments for human degenerative diseases and alleviating the suffering of many human beings."

Immediately afterwards, participants provided their personal judgment of Professor Anderson's action of carrying out this experiment on twenty chimpanzees (*right, wrong,* or *truly divided*). Then, they were asked to explain their reasons for selecting this option, and were provided a large textbox to type their response. After this, they answered two additional questions. They rated on a 1-9 scale how important it was for Professor Anderson to follow the IACUC's ruling (1 = *Not at all important*; 9 = *Extremely important*). They also answered a forced-choice question regarding the chimpanzees' rights to not be harmed: "Do chimpanzees have the right to not be harmed even when the benefits for humankind are significantly great?"

(*Yes, they still have the right not be harmed even when the benefits for humankind are significantly great,* or *No, when the benefits for humankind are significantly great, their rights to not be harmed no longer apply*). Afterwards, all participants were debriefed.

Results

Perceived importance of following authority. There was no difference in participants' ratings of the importance of following authority across the four conditions. Utility definiteness did not affect importance ratings, $F(1, 300) = 1.71$, $p = .415$; authority ruling did not affect importance ratings, $F(1, 300) = 2.56$, $p = .356$; nor did utility and authority interact to affect importance ratings, $F < 1$, $p = .487$. Participants agreed that following the authority of the IACUC was more than moderately important across conditions (means ranged from 6.23 to 6.75), confirming that perceptions of the authority's legitimacy were constant across conditions.

Judgments of the harm. We predicted that authority should exert an influence on judgments predominantly when utility has yet to be obtained. Consistent with this prediction, a multinomial logistic regression of authority and utility definiteness predicting judgments of the harmful act revealed a significant interaction of the independent variables on judgments, $B = 1.45$, $Wald(1) = 4.76$, $p = .029$, when comparing the frequency of "right" judgments with "wrong" judgments. The main effect of authority was also significant, $B = .98$, $Wald(1) = 4.33$, $p = .037$, though the main effect of utility definiteness was not significant, $B = -.64$, $Wald(1) = 1.64$, $p = .20$. Overall, when the authority approved the experiment, fewer participants thought the professor's actions were wrong (14%), and more thought his actions were right (41%), than when the authority disapproved (*wrong* 44%; *right* 21%).

Simple-effects tests confirmed that when utility was merely prospective, a greater percentage of participants judged the scientist's actions wrong when the authority disapproved

(55%) than when it approved (13%), and fewer thought his actions were right when the authority disapproved (18%) than when it approved (49%), $\chi^2(2) = 32.42$, $p < .001$, $\phi_c = .46$ (*truly divided*: 26% vs. 38%, respectively; see Table 3). However, when the utility was definitely obtained, the effect of authority was much smaller and was only marginally significant, $\chi^2(2) = 5.52$, $p = .063$, $\phi_c = .19$; for example, when the utility was definitely obtained, the effects of authority on rightness judgments were 32% (authority rejects) versus 44% (authority approves), as opposed to 18% (authority rejects) deeming the harmful act right versus 49% (authority approves) when the utility was only a prospective possibility (see Table 3).

[Insert Table 3 about here]

Justifications. A coding scheme was developed based on categories adapted from previous research by the authors (see Sousa et al., 2009) and careful examination of participants' responses (see Table 4 for categories and definitions). It is important to note that the definition of most coding categories encompasses two opposing values; for example, the utility category includes both appeals to the utility of the harm (as justification for its rightness) and the lack of sufficient utility (as justification for its wrongness). Almost never did opposing values from a particular category appear within a single justification, though this was theoretically possible. Once the coding scheme was developed, the first author coded all of the responses, and a rater blind to the hypotheses of the study independently coded all of the responses using the coding scheme. Multiple codes were allowed for each case when multiple rationales were presented. Overall, interrater agreement was good (Cohen's $\kappa = .80$); disagreements were resolved via discussion.

[Insert Table 4 about here]

As can be seen in Table 5, participants who judged the harm to be outright "wrong" generally appealed to *justice, rights, and welfare* as justification for their response, though *authority* was a common justification as well, especially within the authority rejects condition. By contrast, participants who judged the harm to be strictly "right" frequently appealed to *utility*. Consistent with our hypothesis, participants who were truly divided tended to invoke both *utility* and *justice, rights, and welfare* arguments.

Rights not to be harmed for human benefits. A chi-square analysis confirmed that perceiving chimpanzees as having the right to not be harmed for human benefit significantly predicted participants judgments, $\chi^2(2) = 109.35$, $p < .001$, $\phi_c = .60$. Participants who perceived that chimpanzees had the right to not to be harmed even for the benefit of humanity were significantly more likely to view the harm as wrong (47%), compared to participants who did not extend this right to chimpanzees (11%). Likewise, only 4% of those who perceived chimpanzees to have this right viewed the professor's actions as right, while 57% of those who believed chimpanzees do not have such rights approved of the professor's actions (*truly divided*: 49% vs. 32%, respectively). Furthermore, participants who endorsed the rights of chimpanzees not to be harmed for human benefit were significantly more likely to justify their decision by appealing to *justice, rights, and welfare* (67%) within their free responses than participants who do not extend these rights to chimpanzees (26%), $\chi^2(2) = 48.99$, $p < .001$, $\phi_c = .40$.

General Discussion

Across three studies, we found that the ruling of an authority significantly affected judgments of utilitarian acts of harm, but this was true mostly when the utility had not yet been realized. In Studies 1 and 3, when a utilitarian act of harm was expected to produce utilitarian benefits, and these benefits had not yet been obtained, participants' judgments of the harm

significantly reflected a concern for whether or not the ruling of an authority had been followed. However, Studies 2-3 clarified that this authority dependence is largely restricted to prospective cases. Harmful acts that have already produced their intended utilitarian benefits elicit much less outright condemnation (i.e., strict "wrong" responses). Nevertheless, many individuals still report being morally divided in these retrospective cases, placing equal weight on the perceived utility of the harm and the rights and welfare of the victim.

These findings help shed light on an unresolved debate within moral psychology regarding the role of authority in judgments of complex harm. Previous research by Kelly et al. (2007) suggests that—contrary to cases of simple harm used by the cognitive-development tradition—there may be complex cases in which people approve the use of harm more when an authority permits it than when an authority prohibits it. However, as discussed in the Introduction, these earlier studies suffered methodological problems that limited the conclusions that could be drawn from them (see also Sousa et al., 2009). In the present studies, we focused on cases of utilitarian harm within a military and scientific context as one class of complex harm (but see Kelly et al., 2007; Sousa et al., 2009; Sousa & Piazza, 2013, for other classes). Rather than simply asking the question of whether or not the consent of an authority would change their response (as in the moral/conventional task used by Turiel and colleagues), or varying the stance of authority in an obvious manner within a within-subjects design (as in both Kelly et al. and Sousa et al.'s studies), we manipulated the ruling of a legitimate authority within a between-subjects design, and afterwards probed participants' understanding of these influences in their judgments using a variety of methods. The advantage of manipulating authority in a between-subjects design is that it circumvents the motivation among participants to be consistent in their judgments. In the context of a within-subjects design, some participants may be reluctant to

change their position regarding the harmful act, for fear of appearing morally capricious, or they may change their judgment to fit the demands of the experimenter. The between-subject design avoids these issues by presenting each participant with only one of the authority conditions. Furthermore, by experimentally manipulating authority, rather than simply probing respondents' views about authority, we were able to determine if authority exerted a causal influence over participants' judgments.

Consistent with previous research using an open-ended format (Sousa et al., 2009), a nontrivial minority of participants in Study 3 reported a concern for authority having been followed as a factor in their judgment, and this was primarily when the authority prohibited the use of the harm (see Table 5). These low rates of authority rationales, however, did not reflect the fairly substantial effect of authority we observed in Studies 1 and 3 (see also Footnote 3). In Study 1, using a more direct assessment, we found that roughly one-third of participants, across the authority conditions, recognized a normative influence of authority on their judgment—a rate more closely approximating the size of the effect we observed in that study. Interestingly, a large percentage of participants continued to report an influence of authority in Study 2 when the effect of authority was softened by utility definiteness. One interpretation of this result is that reports of authority influence actually represent an instance of *post hoc* rationalization (see Haidt, 2001)—in other words, agreement may have occurred only after participants perceived their judgments to be consistent with an authority-based explanation. However, another interpretation of these finding is that for some participants the decree of the authority was a factor that contributed to their conflicted judgment—that is, it kept them from ruling that the act was absolutely the right or wrong thing to do. The fact that most participants who reported an influence of authority in Study 2 were conflicted in their judgment is consistent with this latter

interpretation; nevertheless, future studies are needed to rule out the former possibility. Nevertheless, regardless of whether participants were accurately aware of the role authority played in their judgment, the consistent effect of authority across the present studies leaves little doubt that under conditions of prospective utility, authority does causally influence judgments about inflicting physical harm.

Our findings should not be understood as contradicting the findings of Turiel and colleagues. The focus of this cognitive-developmental tradition has been the study of acts of simple harm, namely, acts performed exclusively for selfish reasons. The current evidence to date supports the notion that acts of simple harm are generally perceived to be wrong regardless of the directives of legitimate authorities, and thus, in the parlance of Turiel and others, qualify as "moral transgressions" (see Sousa & Piazza, 2013). At the same time, our findings suggest that the conceptual criterion of authority independence does not always apply to cases of complex harm (at least not cases of utilitarian harm), since judgments of utilitarian harm were affected to a significant degree by the normative stance of a legitimate authority in some conditions of our studies. Beyond this novel contribution, our findings also help to clarify the parameters under which we might expect complex harm to be authority dependent. Acts of harm with definite utilitarian benefits are easier to justify than when the utility is simply forecasted. When the utility is not in doubt, the relevance of authority ruling appears secondary to the clear utility derived from the act. By contrast, harmful acts that are merely anticipated to bring about utility are somewhat harder to justify given their indefinite status. Thus, when the utilitarian benefits of a harmful act are in question, for a significant number of people, the ruling of an authority aids in adjudicating an otherwise irresolvable moral conflict.

We found that having retrospective information about the utility of a harmful act reduced the influence that authority had on judgments, and generally led to more approval of the harm, compared to cases involving prospective utility. These findings appear at least superficially consistent with studies by Caruso (2010), which found judgments of various "unfair" actions to be less severe when the acts were located in the past versus the near future. There were several differences between our designs that limit direct comparisons: our participants judged the right or wrongness of various acts of utilitarian harm, while Caruso's participants judged the level of fairness of various unfair acts; we manipulated whether *the utility* of a harmful act had or had not yet occurred, while Caruso manipulated whether the unfair act itself had or had not yet occurred; finally, our studies examined the interactive effect of temporal location and authority on judgments, while Caruso's studies focused on the role of negative emotion as a mediator of the effect temporal location had on judgments. Despite these differences, both lines of research seem to be consistent with the general idea that judgments of future events involve some degree of uncertainty and, therefore, are more susceptible to external and internal influences than judgments of past events.

In the present studies, we focused on utilitarian forms of complex harm. However, future research should explore the influence of authority on other classes of complex harm, such as harm committed as just punishment or self-defence—i.e., where the harm fails to impinge on the basic rights of the victim. The studies of Kelly et al. (2007) and Sousa et al. (2009) investigated some of these cases, but found limited support for an influence of authority. Similarly, in an unpublished study, we manipulated between-subjects the position of authority towards harm inflicted on military combatants who gave their consent to be harmed during combat training. We found that most participants approved of the use of harm in this context, irrespective of

authority, since the combatants were aware of the physical risks when they consented to the training. We take these preliminary results as instructive. In cases of complex harm where the victim deserves to be harmed (e.g., harm as punishment) or foregoes their rights to not be harmed (e.g., the person provides their consent to be harmed), or the perpetrator has the right to cause harm (e.g., for self-defence), authority influences are likely to be minimal, due to the core relevance of rights and justice considerations in these judgments, which serve to justify the use of harm in these cases (see Sousa et al., 2009; Sousa & Piazza, 2013, for similar arguments).

Conclusion

We found that many people find the right or wrongness of utilitarian harm to be dependent on the normative position of an authority. However, this seems primarily to hold when the forecasted benefits of the harm have not yet been obtained. Thus, although attitudes towards utilitarian harm vary widely, we would expect judgments to shift most dramatically when an authority endorses or forbids an act of harm that has the *potential* for utilitarian benefits. Conversely, our findings suggest that the best way to attenuate authority influence in such situations is to educate individuals about the definite utility or non-utility of particular harmful acts—a tactic that seems increasingly tenable in debates over the use of torture interrogation (Brumfiel, 2006; Janoff-Bulman, 2007) and animal testing (Bekoff, 2007; Singer, 2002).

Acknowledgements

Thanks to Jennifer Lord, Catherine Peralta, and Jillian Tuso for their assistance with these studies, and three anonymous reviewers for their invaluable feedback on an earlier version of this paper.

References

Bekoff, M. (2007). *Animals matter: A biologist explains why we should treat animals with compassion and respect.* Boston, MA: Shambhala.

Brumfiel, G. (2007). Interrogation comes under fire. *Nature, 445(25),* 349

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6,* 3-5.

Caruso, E. M. (2010). When the future feels worse than the past: A temporal inconsistency in moral judgment. *Journal of Experimental Psychology: General, 139,* 610-624.

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,* 1082-1089.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review, 5,* 5-15.

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.

Gert, B. (2004). *Common morality: Deciding what to do.* New York: Oxford University Press.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44,* 389-400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293,* 2105-2108.

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

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

Janoff-Bulman, R. (2007). Erroneous assumptions: Popular belief in the effectiveness of torture interrogations. *Journal of Peace Psychology, 13,* 429-435.

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

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature, 446,* 908-911.

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. A.Goslin (Ed.), *Handbook of socialization theory and research* (pp. 347-480). Chicago, IL: RandMcNally.

Laupa, M., & Turiel, E. (1986). Children's conceptions of adult and peer authority. *Child Development, 57,* 405-412.

Laupa, M., & Turiel, E. (1993). Children's concepts of authority and social contexts. *Journal of Educational Psychology, 85,* 191-197.

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition, 100,* 530-542.

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.

Nucci, L., & Turiel, E. (1993). God's word, religious rules, and their relation to Christian and

Jewish children's concepts of morality. *Child Development, 64,* 1475-1491.

Singer, P. (2002). *Animal liberation.* New York: HarperCollins.

Singer, P., & Cavalieri, P. (Eds.). (1993). *The Great Ape Project; Equality beyond humanity.* London: Fourth Estate Publishing.

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

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

Sousa, P., Holbrook, C., & Piazza, J. (2009). The morality of harm. *Cognition, 113,* 80-92.

Sousa, P., & Piazza, J. (2013). *Harmful transgressions qua moral transgressions: A deflationary view.* Manuscript submitted for publication.

Stich, S., Fessler, D. M. T., & Kelly, D. (2009). On the Morality of Harm: A response to Sousa, Holbrook, and Piazza. *Cognition, 113,* 93-97.

Thomson, J. J. (1985). The trolley problem. *Yale Law Journal, 94,* 1395-1415.

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., Hildebrandt, C., & Wainryb, C. (1991). Judging social issues: Difficulties,

inconsistencies and consistencies. *Monographs for the Society of Research in Child Development 56* (Serial no. 224).

Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science, 17,* 476-477.

Wadman, M. (2012, July 25). Bill to end US chimp research advances. *Nature.com.* Retrieved from http://blogs.nature.com/news/2012/07/bill-ending-us-chimp-research-advances.html

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,* 780-784.

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

Table 1

Frequency of judgment as a function of authority condition (Study 1)

	Authority	
	Prescribe	Prohibit
Right	45	39
Truly divided	30	23
Wrong	7	19
Total	82	81

Table 2

Frequency of judgment as a function of authority condition: Definite utility derived from the harm (Study 2)

	Authority	
	Prescribe	Prohibit
Right	28	28
Truly divided	31	28
Wrong	7	12
Total	66	68

Table 3

Judgments of the utilitarian harm from Study 3 as a function of utility definiteness and authority ruling

	Definite Utility		Indefinite Utility	
	Approves	Rejects	Approves	Rejects
Right	24	18	**38**	14
Truly Divided	**40**	**33**	29	20
Wrong	12	24	10	**42**
Total	76	75	77	76

Note. Values in bold represent the modal response for that condition.

Table 4

Coding scheme: Justification categories and definitions used in Study 3

Category	Definition
Justice, rights and welfare (JRW)	Appeals to the harm or suffering caused to the target or lack of harm or suffering caused, or the justice or injustice of the act; this includes considerations of the degree of harm, the innocence of the target, the targets' right not to be harmed, or whether the target deserved or did not deserve to be harmed in such a manner.
Utility	Appeals to the utility of the act, or the utility expected of the act, as a means for promoting a greater good (e.g., alleviating human suffering); or an appeal to the act not providing sufficient utility to justify the harm done.
Moral value	Appeals to the moral value of chimpanzees (e.g., due to their intelligence), or their lesser value compared to humans.
Authority	Appeals to the fact that an authority (the IACUC) approved or rejected the act, or a concern that the authority was followed or not followed.
Unscorable/Restatement	Participant's reasoning is unclear or he/she simply restates their judgment (e.g., that the act was wrong) without further elaboration. Used only when no other category applied.

Table 5

Justifications as a function of utility, authority, and judgment (Study 3)

	Definite Utility						Indefinite Utility					
	Approves			Rejects			Approves			Rejects		
	Right	Divided	Wrong	Right	Divided	Wrong	Right	Divided	Wrong	Right	Divided	Wrong
JRW	1	27	**11**	3	18	12	4	22	**8**	1	10	**24**
Utility	**23**	**31**	0	**13**	**30**	2	**34**	**23**	0	**12**	**16**	0
Moral value	3	6	6	8	2	2	6	4	2	3	2	7
Authority	2	2	0	0	12	**14**	3	1	0	0	4	22
Unscorable	0	2	0	0	1	3	1	1	1	1	0	1

Note. JRW = justice, rights, and welfare. Unscorable includes restatements. Values in bold represent the modal justification for judging the harm as right, truly divided, or wrong within each condition.