Simulating murder: The aversion to harmful action

Fiery Cushman[1], Kurt Gray[2], Allison Gaffey[3] & Wendy Berry Mendes[4]
[1]Harvard University, [2]The University of Maryland, [3]University of Notre Dame [4]University of California, San Francisco

Diverse lines of evidence point to a basic human aversion to physically harming others. First, we demonstrate that unwillingness to endorse harm in a moral dilemma is predicted by individual differences in aversive reactivity, as indexed by peripheral vasoconstriction. Next, we tested the specific factors that elicit the aversive response to harm. Participants performed actions such as discharging a fake gun into the face of the experimenter, fully informed that the actions were pretend and harmless. These simulated harmful actions increased peripheral vasoconstriction significantly more than did witnessing pretend harmful actions or to performing metabolically-matched non-harmful actions. This suggests that the aversion to harmful actions extends beyond empathic concern for victim harm. Together, these studies demonstrate a link between the body and moral decision making processes.

People are averse to performing harmful actions and often consider it morally wrong to harm a person even when it would save many more lives (Mikhail, 2000; Petrinovich, O'Neill, & Jorgensen, 1993). Even front-line soldiers trained and motivated to kill often deliberately miss visible enemy targets (Grossman, 1995). This aversion to harm is essential to ordinary human functioning, as evidenced by the antisocial behavior of psychopaths, who are argued to lack it (Blair, 1995). The aversion to harming others is so basic to our moral sense that it is easy to miss an important question: What is its psychological basis?

Our first experiment examines the link between physiological responses and answers on a classic moral dilemma: whether it is allowable to kill someone in order to save many lives. We measured physiological reactivity, linked to general aversive states, during a non-moral task, and then examined whether it predicted advocating the death of one person to save others. We use autonomic changes, specifically changes in total peripheral resistance (TPR), which are associated with negative stress responses (Gregg, James, Matyas, & Thorsteinsson, 1999; Mendes, Blascovich, Hunter, Lickel, & Jost, 2007). Past research indicates that the aversion to harm in such moral dilemmas involves an affective component (e.g. Greene, Sommerville, Nystrom, Darley, & Cohen, 2001, Koenigs et al., 2007; Mendez et al., 2005, Moretto, Làdavas, Mattioli, & di Pellegrino, 2010), but does not establish a link between the specific aversion to harm in moral dilemmas and general aversive reactivity in non-moral situations.

Our second experiment builds on this finding, asking why people find the performance of harmful actions aversive. First, aversion may stem from empathic concern for the welfare of the victim (Crockett et al., 2010; Hoffman, 2000; Mehrabian & Epstein, 1972; Pizarro, 2000). For instance, we might be averse to punching another because considering the victim's pain causes us psychological distress (Batson et al., 2003; Gray, Gray & Wegner, 2007; Singer, et al., 2004). Importantly, victim distress is not an intrinsic property of an *action* itself, but rather of its expected *outcome*. We call this the "outcome aversion" model: people are averse to harmful acts because of empathic concern for victim distress.

In addition, an aversive response might be triggered by the basic perceptual and motoric properties of an action, even without considering its outcome. Blair (1995) suggests a mechanism by which harmful actions themselves can become aversive: When the unconditioned aversive stimulus of victim distress (e.g. crying) is repeatedly paired with a particular action (e.g. pushing or hitting a person), those actions acquire a conditioned aversive response. On this "action aversion" model, empathy is critical to the acquisition of the aversive response to harmful actions, but the conditioned response may subsequently be evoked by intrinsic properties of the action alone.

Study 2 tests for "action aversion" by examining participants' physiological responses while either performing or witnessing harmful actions (e.g., stabbing an experimenter with a rubber knife, shooting him with a disabled handgun, etc.), or performing similar but harmless actions (e.g., slicing a pretend loaf of bread with a knife). Our use of simulated actions follows past research demonstrating that pretend stimuli can be sufficient to elicit strong psychological responses (Rozin, Millman, &

Nemeroff, 1986). Action aversion predicts a robust aversive response to pretend actions with motoric and perceptual properties of actual harmful behaviors even though the "perpetrator" knows that no harm will occur, whereas outcome aversion does not. Additionally, action aversion predicts a greater aversive response to performing harm than witnessing it (because only the former involves an action), whereas outcome aversion predicts an equal aversive response in both cases (because they yield the same outcome).

In summary, in Study 1 we test the relationship between the moral judgment of harmful actions and general TPR reactivity to a non-moral task. In Study 2 we test whether simulated harms specifically trigger TPR reactivity. Additionally, we compare reactivity for performing versus witnessing simulated harm, testing whether reactivity depends on the anticipation of a harmful outcome versus the performance of a harmful act.

Study 1

Study 1 examines the relationship between threat reactivity and responses to a classic moral dilemma. We tracked changes in TPR during a stressful arithmetic task and predicted that individuals exhibiting greater TPR reactivity would be less willing to endorse harming one person in order to save the lives of several others.

Method

We recruited 108 healthy participants (81 female) aged 19-40 years (median 24). After obtaining consent, an experimenter applied sensors that measured impedance cardiography (HIC 2500, Chapel Hill, NC)), electrocardiography (Biopac ECG module, Goleta, CA), and blood pressure responses (Colin Prodigy II, San Antonio, TX). Impedance cardiographic and electrocardiography signals were sampled at 1000Hz and integrated with a Biopac MP150. Post-acquisition waveforms were scored using Mindware software (IMP 3.0) by trained research assistants (see Mendes, 2009). We estimated TPR using the standard formula:

TPR = (Mean arterial pressure/Cardiac output) x 80.

After baseline participants met a new experimenter who asked them to count backwards quickly in steps of 7 from a four digit number. Mental arithmetic is a common laboratory stress task that can evoke increases in sympathetic nervous system responding. We used TPR change during the first minute of the stress task as our indication of threat reactivity, subtracting the last minute of the baseline period from the first minute of the stress task.

Participants were recruited as part of a larger study on physiological changes associated with emotion and body manipulations, analyses of which are beyond the scope of the present study. Here we report physiological responses that occur prior to our emotion manipulations. The body position manipulation (leaning forward versus leaning back) was introduced before the stress task, but had no effect on the early physiological responses we analyze here.

Participants were then provided a packet of questionnaires that included a moral dilemma. It asked the participant to imagine being on a lifeboat that would sink – killing all onboard –unless someone is thrown off. One person on the lifeboat is "leaning over the side." Participants were asked, "Is it morally acceptable for you to push this person overboard in order to save the lives of the remaining passengers", indicating yes/no. Then they were asked, "Please indicate how morally acceptable it would be for you to throw this person overboard in order to save the lives of the remaining passengers" on a seven point scale (1 = "completely unacceptable", 7= "completely acceptable").

Results and Discussion

Measurement of TPR requires several high-quality, artifact-free physiological signals. Twenty participants were excluded for low-quality impedance or electrocardiograph waveform, twenty-three because blood pressure measurements were not obtained during the first minute of the task, three because TPR reactivity scores differed by more than two standard deviations from the mean, and another nine dropped out of the study prior to the assessment of the moral dilemmas. Responses to moral dilemmas did not significantly differ between those with usable TPR data to those without, t (106) = 0.60, *ns*.

As predicted, increased TPR reactivity[2] was reliably associated with lesser endorsement of pushing a person overboard in our moral judgment task r = -.31, N = 51, $p < .05$, although not when rated dichotomously, $t(40) < .01$, *ns*. This result was robust after controlling for the experimental manipulations of posture, affect, their interaction, gender, and age, r = -.31, N = 51, $p < .05$.

The observed correlation between moral judgment and TPR reactivity is consistent with our

prediction that unwillingness to endorse harmful action is linked with threat reactivity. This suggests that the aversion to harmful actions may be instantiated physiologically. However, it leaves open the basis of this response: Does it depend upon empathy for an actual victim, or also upon the perceptual and motoric properties of the action itself?

Study 2

Study 2 tested the action aversion hypothesis – whether physiological aversion can be triggered by only the motor or perceptual properties of harmful action. Participants were asked to perform five simulated harmful actions, to witness another person perform them, or to perform five simulated non-harmful actions. We tested two predictions of the action aversion hypothesis: simulated harmful actions would elicit aversive reactivity despite the absence of any harmful outcome, and this response would be greater when performing the action than when witnessing the action.

Method

We recruited 108 participants (69 female) aged 18-35 years (median 20). Participants initially consented to a study of "pretend actions" omitting any mention of harm. Participants completed twenty items from the Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988) and then we applied sensors from two different ambulatory devices that allowed free movement during testing: VU-AMS impedance/ECG (Amsterdam, The Netherlands), and SpaceLab blood pressure (ABP 90207, Issaquah, WA). The experimenter measured baseline physiological responses during 5min rest, and then brought participants into a room where she described the full experimental procedure, obtained additional informed consent, and initiated the pretend actions. The experimenter emphasized that the participant was free to omit any actions.

A male research assistant played the role of victim in the "perform harm" and "witness harm" conditions. Participants in the "perform harm" condition were asked to perform five actions in a fixed order: (1) smashing the victim's shin with a hammer – a PVC pipe was worn under a fake pant leg, (2) smashing the victim's hand with a rock – a rubber hand was placed at the cuff of the shirt and the actual hand was obscured from sight, (3) discharging a handgun into the victim's face – a weighty metal replica, (4) drawing a knife across the victim's throat – a rubber knife, and (5) smacking a baby against the table – we used a realistic looking baby doll (see Figure 1). No verbal communication occurred between the participant and the victim, and the victim avoided eye contact except during the action itself. The victim grimaced slightly during each action, but exhibited no further distress.

The experimenter initiated each action by describing it to the subject and emphasizing that it was pretend and no harm would occur. The participant was instructed to contemplate performing the action for one minute while holding the relevant implement (e.g. gun), during which the experimenter exited the room. The experimenter returned and said, "It is time to perform the action." After the subject performed the action (or chose not to) the experimenter closed a curtain between the participant and the victim, instructed the participant to sit quietly for one minute, and exited. This sequence was repeated for each action.

The "witness harm" condition proceeded identically except that the participant was introduced to two additional experimenters who assumed the roles of perpetrator and victim. The participant heard an identical description of the event to be performed, and then contemplated watching that event for one minute. The experimenter returned to the room and asked the perpetrator to "harm" the victim, which the perpetrator did directly in front of the subject with neutral affect. The perpetrator and the victim were then masked by a curtain during a one minute post-action period. The "no harm" condition also proceeded identically except that there were no additional experimenters, and the participant was asked to perform five metabolically-controlled pretend actions: (1) hammering an imaginary nail on a block of wood, (2) using a rock to smash a (rubber) nut, (3) using a spray bottle to mist an imaginary plant, (4) using a rubber knife to cut a (cardboard) loaf of bread, and (5) smacking a hand broom against a table to shake out dust.

After the pretend actions, participants returned to the original room and sensors were removed. Participants completed several questionnaires including the 20-item PANAS and 5 hypothetical moral dilemmas drawn from previous research (Greene et al., 2001). These moral dilemmas asked participants whether they would perform a harmful action in order to save many lives (e.g. whether to smother one's own crying baby in order to successfully hide the whole family from enemy soldiers).

During the experiment impedance and electrocardiograph were monitored continuously, but BP measurements were manually initiated by the experimenter with a key press. We took BP readings at the first, third and fifth actions, and the timing of these (before versus after the action) was varied between participants. Thus, for each subject we calculated three TPR reactivity scores.

One subject elected not to perform any actions and was excluded from all analyses. Thirty-three participants were excluded from physiological analyses due to equipment malfunction, experimenter error, or biologically implausible measurements. Individual data points greater than two standard deviations from the group mean were excluded from analysis. Several participants were excluded from behavioral analysis because of incomplete responses to post-test survey items.

Results

Self-reported affect

An ANCOVA of post task negative affect (controlling for pre-task scores) yielded a significantly effect $F(2, 101) = 16.90$, $p < .001$. Simple effects revealed increased negative affect for perform harm ($p < .001$) and witness harm ($p < .001$) compared with no harm, but no difference between perform and witness conditions ($p > .25$).

Harmful Actions

We used the general linear model to test for effects of condition (perform vs. witness vs. no harm) and measurement period (pre-action contemplation vs. post-action recovery) on TPR reactivity, treating participant as a random effect and employing robust standard errors. There was a significant effect of measurement period $\beta_1 = -.31$, $z = -3.30$, $p<.01$: Across all three conditions, TPR reactivity was greater pre-action than post-action. We treated the perform harm condition as the comparison condition, and found that TPR reactivity was significantly greater than for witness harm $\beta_2 = -.23$, $z = -2.08$, $p<.05$ and no harm $\beta_3 = -.26$, $z = -2.65$, $p<.01$ (Table 1). Additional analyses revealed no significant effect of order (earlier versus later actions) on TPR reactivity, and no significant interactions between condition, measurement period, or task order.

We performed a supplementary analysis of TPR reactivity taken from the very *first* pre-action contemplation period, a point at which participants had anticipated the task, but had not performed or witnessed any action. TPR reactivity for the perform condition ($M=89.8$) was significantly higher than for the witness condition ($M=30.7$) $\beta = -.37$, $t=2.36$, $p<.05$ and the no harm condition ($M=29.9$) $\beta=-.37$, $t=2.34$, $p<.05$.

Moral Dilemmas

Paralleling our analysis in Study 1, we assessed whether greater TPR reactivity was associated with lesser endorsement of harming one person in order to save several others. We calculated a summary TPR score for each participant, averaging their reactivity measurements and adjusting post-action TPR scores to match pre-action TPR scores according to the relevant coefficient (β_1) of the GLM presented above. We then correlated TPR reactivity with the mean judgment across five moral dilemmas. Collapsed across all three conditions this correlation was significant $r = -.32$, $N=71$ $p<.01$. The relationship was larger and significant for the witness harm condition $r = -.49$, 95% CI -.12- -.87, $N=25$, $p<.05$, smaller and non-significant correlation for the perform harm condition $r = -.36$, CI -.06-.78, $N=23$, $p<.10$, and smallest and non-significant for the no harm condition $r = -.24$, CI -.20-.68, $N=23$, $p<.28$. Analysis of variance revealed no significant effect of condition on mean moral judgment $F(2, 97)=0.75$, $p=.47$. Even though our condition effects did not influence later moral judgments on hypothetical scenarios, individual differences in TPR reactivity predicted moral judgments as in Study 1.

General Discussion

We investigated individuals' aversion to harmful actions. Study 1 demonstrated that individuals exhibiting greater threat reactivity were less likely to endorse harm in order to save lives. This finding corroborates past evidence suggesting a more potent aversive response to the idea of performing direct harm than to allowing indirect harm to more distant others. It also suggests that individual differences in total peripheral resistance during non-stressful tasks can predict this aversion to harm.

Study 2 investigated the psychological basis of this aversive response. Performing simulated harmful actions evoked robust TPR reactivity despite participants' full awareness that no actual harm would be caused. TPR reactivity was lower among participants asked to witness harmful actions or to perform metabolically-controlled non-harmful actions. Moreover, TPR reactivity differed between conditions

during the very first pre-action measurement period, prior to participants performing or witnessing any action at all. Thus, simply contemplating performing a simulated harmful action leads to greater vasoconstriction than contemplating witnessing a harmful action.

These findings suggest an aversion to performing harmful actions that extends beyond the expectation of a harmful outcome. Clearly, TPR reactivity to *pretend* harmful action (such as hitting a plastic baby doll) cannot be attributed to explicit belief that harm will occur. Nevertheless, pretend events could trigger the imagination of harmful outcomes. Critically, however, outcome aversion predicts similar affective states for witnessing and performing harm, while action aversion predicts a unique aversive response to performing harm, as we observed. We therefore consider it unlikely that the TPR reactivity associated with performing simulated harmful actions was caused solely by consideration of a harmful outcome, such as empathic concern for victim distress.

Our findings do not contradict the role of empathy and victim distress on the aversion to harmful actions. These elements may be especially important for the performance of "real" harmful actions with actual consequences; moreover, they may play a key developmental acquisition of action aversion via associative pairing (Blair, 1995). Just as you cannot help but swoon when smelling the perfume or cologne associated with your first love, people cannot help but feel upset when doing actions typically associated with victim distress. Yet, as important as the aversion to victim distress may be, our results suggest a dissociable aversion based on mere actions.

A forceful, automatic aversive response to the surface properties of harmful actions may explain otherwise puzzling human behaviors. In battlefield behavior and hypothetical moral judgment people resist doing direct harm despite explicit knowledge that it could save many lives. Similarly, in our study, people experienced a strong aversive response to performing pretend harmful actions despite the explicit knowledge that no harm would be caused. These cases highlight a dissociation between our explicit knowledge of the consequences of our actions and our automatic affective responses to actions (Dayan & Niv, 2008; Kahneman, 2003; LeDoux, 1996).

The action aversion model also suggests a darker side: when banal or novel actions lack motoric and perceptual properties associated with harm, they may fail to trigger an aversive response. Signing one's name to a torture order or pressing the button that releases a bomb each have real, known consequences for other people, but as actions they lack salient properties reliably associated with victim distress. A notable parallel is evident in moral judgment: People consider it morally worse to cause harm through direct physical engagement than at a distance (Cushman, Young & Hauser, 2006; Greene, Cushman, Stewart, Lowenberg, Nystrom & Cohen, 2009). We demonstrate that TPR reactivity increases during (pretend) harmful actions and also correlates with judgments of moral dilemmas. Yet, while circumstantial evidence implicates a role for action aversion in moral judgment, further research is required.

As such, our study highlights the advantage of taking lessons from hypothetical moral dilemmas and translating them into more active behaviors. Few past studies directly target the human aversion to harm using an active behavioral paradigm (Martens, Kosloff, Greenberg, Landau, & Schmader, 2007; Milgram, 1974). This is no surprise: it is hard to get one person to harm another ethically and in a laboratory. Moreover, past studies often targeted situational factors that promote harm, rather than the affective systems that discourage it. To ask why people *do* harm is a critical research question; our complementary question is why people *do not*. Our study suggests that the use of simulated harmful actions is sufficient to generate an aversive response. Surely this response is weaker than the aversion experienced by a solider on the battlefield, or the captain of a sinking ship; nevertheless, the aversion to simulated harm in the laboratory may provide insight into the psychology underlying the aversion to actual harm in the world beyond.

Acknowledgements

We thank our undergraduate research assistants and other members of the Health and Physiology Laboratory and Moral Cognition Laboratory for their assistance in collecting the data. This research was supported by funding from NHLBI grant (HL079383) and the Mind, Brain and Behavior Initiative at Harvard University.

References

Batson, D.C., Lishner, D.A., Carpenter, A., Dulin, L., Harjusola-Webb, S., Stocks, E.L., Gale, S., Hassan, O. & Sampat, B. (1993). "…As you would have them do onto you": Does imagining yourself in the other's place stimulate moral action? *Personality and Social Psychology Bulletin*, *29*(9), 1190-1201.

Blair, R. J. R. (1995). A cognitive developmental approach to morality: investigating the psychopath. *Cognition, 57*, 1-29.

Cushman, F. A., Young, L., & Hauser, M. D. (2006). The role of conscious reasoning and intuitions in moral judgment: Testing three principles of harm. Psychological Science, 17(12), 1082-1089.

Crockett, M., Clark, L., Hauser, M., & Robbins, T. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences, 107*(40), 17433.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology, 18*, 185-196.

Gray H.M., Gray K., & Wegner D.M. (2007). Dimensions of mind perception. *Science*, 315(5812), 619.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. Cognition, 111(3), 364-371.

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.

Grossman, D. (1995). *On killing*. Boston: Little, Brown and Company.

Gregg, M. E., James, J. E., Matyas, T. A., & Thorsteinsson, E. B. (1999). Hemodynamic profile of stress-induced anticipation and recovery. International Journal of Psychophysiology, 34, 147-162.

Hoffman, M. L. (2000). *Empathy and Moral Development*. Cambridge: Cambridge University Press.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American economic review, 93*(5), 1449-1475.

LeDoux, J. (1996). *The Emotional Brain*. New York: Simon and Shuster.

Martens, A., Kosloff, S., Greenberg, J., Landau, M., & Schmader, T. (2007). Killing begets killing: Evidence from a bug-killing paradigm that initial killing fuels subsequent killing. *Personality and Social Psychology Bulletin, 33*(9), 1251.

Mehrabian, A., & Epstein, N. (1972). A measure of emotional empathy. *Journal of Personality, 40*(4), 525-543.

Mendes, W., Blascovich, J., Hunter, S., Lickel, B., & Jost, J. (2007). Threatened by the unexpected: Physiological responses during social interactions with expectancy-violating partners. *Journal of Personality and Social Psychology, 92*(4), 698.

Mikhail, J. M. (2000). *Rawls' Linguistic Analogy: A study of the 'generative grammar' model of moral theory described by John Rawls in 'A theory of justice'*. Unpublished Doctoral Dissertation, Cornell University, Ithaca.

Milgram, S. (1974). *Obedience to Authority: An experimental view*. New York: Harper & Row Publishers, Inc.

Moretto, G., Lǂdavas, E., Mattioli, F., & di Pellegrino, G. (2010). A psychophysiological investigation of moral judgment after ventromedial prefrontal damage. *Journal of Cognitive Neuroscience, 22*(8), 1888-1899.

Petrinovich, L., O'Neill, P., & Jorgensen, M. J. (1993). An empirical study of moral intuitions: towards an evolutionary ethics. *Journal of Personality and Social Psychology, 64*(3), 467-478.

Pizarro, D. A. (2000). Nothing more than feelings? The role of emotions in moral judgment. *Journal for the Theory of Social Behavior, 30*(4), 355-375.

Rozin, P., Millman, L., & Nemeroff, C. (1986). Operation of the laws of sympathetic magic in disgust and other domains. *Journal of Personality and Social Psychology, 50*, 703-712.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, 54(6), 1063-1070.

Footnotes

1. We also examined the association between responses on moral dilemmas and changes in cardiac output. Consistent with the "threat" profile CO decreases were associated with less endorsement of harming others, r (N=54) = .29, $p < .04$. In Study 2 we examine responses during a task that does not meet the requirements of a motivated performance situation, so we focus on TPR in both studies for consistency.

Figure 1: Harmful and non-harmful actions used in Experiment 2.

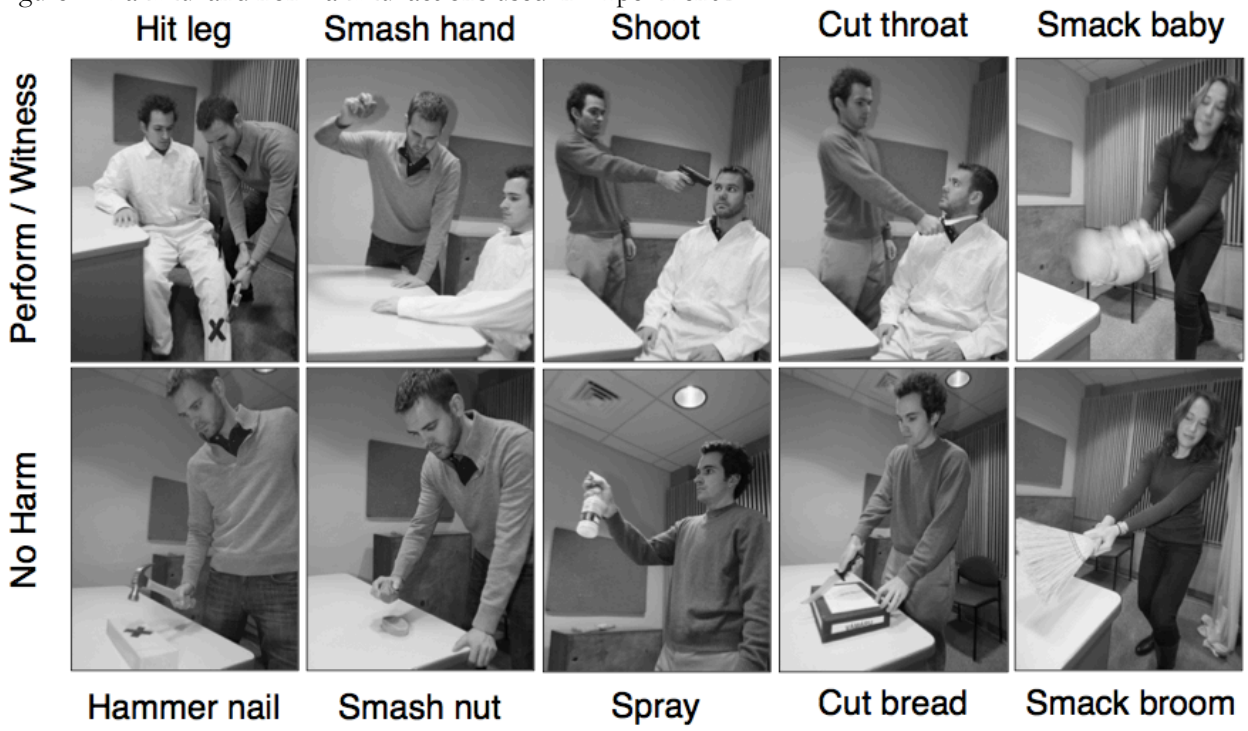

Table 1: Mean TPR reactivity by condition and measurement period. N indicates number of observations, with up to three observations per subject

	TPR Change: M *(SD, N)*	
Condition	*Pre-action*	*Post-action*
Perform	75 *(73, 34)*	47 *(90, 28)*
Witness	34 *(73, 49)*	10 *(88, 24)*
No Harm	29 *(83, 48)*	-28 *(68, 16)*