Acknowledging and Managing Deep Constraints on Moral Agency and the Self: Comment on Doris (2017)

Laura Niemi
Department of Psychology, Harvard University
lauraniemi@fas.harvard.edu

Jesse Graham
Department of Psychology, University of Southern California
jesse.graham@usc.edu

in press, *Behavioral and Brain Sciences*

Abstract
Doris proposes that the exercise of morally responsible agency unfolds as a collaborative dialogue among selves expressing their values while being subject to ever-present constraints. We assess the fit of Doris' account with recent data from psychology and neuroscience related to how people make judgments about moral agency (responsibility, blame), and how they understand the self after traumatic events.

In *Talking to Our Selves*, Doris (2015) grapples with the problem of whether and how people ought to be considered morally responsible agents when they do not seem to be able to access accurate accounts of the reasons for their own behaviors. He spends a good portion of the book gathering findings from psychology experiments to demonstrate that people are better at fooling themselves than knowing themselves. We act with a number of arbitrary and ridiculous influences pulling the strings, and when we attempt to explain ourselves by looking inside, we wear rose-colored, and awfully smudged, glasses. Thus, taking psychological science seriously, Doris positions himself as skeptical about people's ability to exercise morally responsible agency. However, Doris contends that people may sometimes exercise morally responsible agency -- to the extent that their behaviors express their values. How could we possibly know when a person's behavior expresses his or her values? Given rampant self-deception and self-ignorance, we're asked to be wary of what probably strikes as a good signal: a person's willingness to mobilize a verbal defense of his or her behaviors. The *dialogic* or *collaborativist* aspect of Doris' theory addresses worries about how to precisely determine when a person has acted according to his or her values, by pointing out that the continual cognitive penetration of people's evaluative judgments by external forces -- including, importantly, the value-driven

questioning of others that occurs in dialogue -- renders their values not truly their *own*. The collaborativist view of agency hinges on a collaborativist notion of the *self* in which what individuals count as valuable for the self depends on what other people count as valuable.

Moral responsibility, in turn, is determined through exchange and negotiation of reasons, in an unfolding, collaborative conversation. Ostensibly, as happens in negotiations, for a matter to be considered settled on both ends, both parties will trade off pleading and conceding until they can peacefully move on from it. So, it is more than okay to consider people self-directed value-driven agents when they, for example, initially claim ignorance about moral permissibility or when they are unable to articulate their position, in addition to easier cases such as when they appear to be squeamish about making value judgments or taking a stand. By equating agency with negotiation, a collaborativist view of moral agency "trades in uncertainty" (p. 13, PRÉCIS); and is normatively neutral. Interestingly, one way Doris' account of the exercise of moral agency maintains neutrality is that it accommodates an interpretation that is congenial with people's interests in social justice (moral agency is participatory action), but also maintains throughout a deeper, sometimes unsettling, message about constraints (moral agency is inevitably never truly up to one person).

As psychologists pursuing the scientific study of the unruly domain of morality, we consider Doris' empirically-based philosophy of moral agency an endlessly thought-provoking accomplishment. In the spirit of collaborative conversation, we offer up some more data from psychological science and assess how it places within his account.

Doris' account of the exercise of morally responsible agency is dialogic but largely focuses on the exercise from one side, the perspective of the doer. What about the other side, the observer or judge? How do people go about making judgments about *others* relevant to morally responsible agency? First, surveys of people across the globe over the last decade allow us to be more certain about what people explicitly value. Namely, there is solid evidence that caring and compassion are broadly valued, whereas harm and exploitation are inconsistent with most people's values (e.g., Haidt, 2007; Graham et al., 2011). This suggests that on the aggregate people should not be "victim blamers" -- they should attribute blame and responsibility so that

harm-doers do not get off the hook, and vulnerable people who have been harmed are protected. In our own vignette studies, this is largely how participants make judgments: people (who were not explicitly labeled "perpetrators") who robbed and sexually assaulted were attributed more responsibility and blame than those who were robbed or assaulted (who were not explicitly labeled "victims"), and people higher in caring values rated explicitly labeled victims more injured and wounded (Niemi & Young, 2016). These findings are consistent with findings from moral psychology that demonstrate people's general aversion to directly harming people and their weighting of information about kindness and compassion in person perception (e.g., Greene, Sommerville, Nystrom, Darley & Cohen, 2001; Goodwin, 2015; Miller, Hannikainen & Cushman, 2014). Recent neuroscientific work links moral condemnation of harm to normally functioning emotional processing (e.g., Crockett, Clark, Hauser, Robbins, 2010; Greene et al., 2001; Koenigs et al., 2008; Park, Kappes, Rho, Van Bavel, 2016; Perkins et al., 2014). Taken together, these findings indicate equating agency with the term *negotiation* doesn't fit with how people go about moral judgment in cases of direct inducement of bodily harm. In these cases, agency isn't negotiated; it probably never makes it close to the negotiation table because most people's biology supports values that reflect concern about bodies as protected from painful imposition.

In an approach complementing Doris' "ecumenical pluralism" (Doris, p. 186) with respect to agency, continuity, and identity, these massive survey efforts took a moral pluralist approach, going beyond the values of WEIRD participants. Findings revealed not only broad shared valuation of care, but also variability in people's endorsement of statements reflecting the values of loyalty, obedience to authority, and concern about purity (*binding values;* Graham, Haidt & Nosek, 2009; Graham et al., 2011). Strikingly, some people explicitly rank concern about binding values equivalently to concern about "doing no harm," whereas others seem offended by the very idea of such a prospect. People higher in binding values tend to also endorse higher levels of religiosity and political conservatism. We have found that the more people endorse the binding values of loyalty, obedience to authority, and concern about purity (controlling for gender, politics and religiosity), the more they appear like "victim blamers" -- they are more likely to attribute blame and responsibility to victims, say a change in the victim's' actions would have made a difference to the outcome, rate victims as contaminated and tainted,

and generate fewer counterfactual statements about perpetrator behavior when asked "how could the outcome have been different" (Niemi & Young, 2016).

These findings indicate that, in addition to amending Doris' *valuational* theory of moral agency to account for the role of the body and more broadly shared valuation of compassionate caring (i.e., that biologically-based aversion to harm allows for some reasonable predictions about action and moral judgment), the proposition that values are central motivators of action is underspecified in another way: modern culture may be unified about caring, but it's not unified about loyalty, obedience or purity. And explicit endorsement of binding values is reliably related to how people attribute responsibility and blame across the moral dyad of agent and patient. That is, the extent to which people value obligations at a more abstract level related to loyalty, obedience, and purity relates to how much they factor the contributions of *affected* individuals -- moral patients -- into their moral judgments. These judgments of responsibility, blame, and contamination have the potential to be consequential to individuals' well-being and personal freedom.

However, consistent with Doris's account, people's judgments were still also influenced by factors outside their awareness. We experimentally manipulated linguistic focus on agents versus patients in vignettes involving sexual assault by placing the perpetrator (agent) in the subject position in the majority of sentences for half the participants, and the victim (patient) in the subject position for the other half. When people focused on victims (patients), they attributed them more responsibility and blame compared to when they focused on perpetrators (agents) -- this implicit influence factored into ratings of responsibility and blame in addition to binding values (Niemi & Young, 2016). These findings may be taken as some evidence that *judgment* of morally responsible agency across the agent-patient dyad can unfold similarly to how Doris proposes moral agency unfolds from the first-person perspective: as an exercise of values penetrated by implicit influences.

What can psychological science say about how values might relate to perception of the self? In the last chapter of the book, Doris expands on the notion of the *socially contingent self*, crucial to his collaborativist view of moral agency. To do so, he shifts from how individuals are

constrained even when they feel their most *able*, to a complementary and illuminating theme: how the severing of meaningful social ties apparently leaves individuals feeling completely *disabled.* In a striking passage, Doris describes how the last-surviving members of the Crow tribe, subjected to cultural annihilation, subsequently reported existential emptiness -- as though they had "predeceased their bodies." Doris contends that cultural devastation experienced by members of the tribe led to a specific kind of intra-psychological change: rupture in the sense of continuity of the self, as though they were "no longer the same person" (Doris, p. 183).

Do people really endorse disruptions of personal continuity like this? Indeed they do. Trauma-related cognitions including beliefs about a *foreshortened future* align with the self-rupture Doris describes -- e.g., "My life has been destroyed." "I feel like I don't know myself anymore." "I've lost my soul forever." "I feel dead inside." "My life will never be the same again." (Ehlers & Clark, 2000; Foa et al., 1999; Niemi & Nizzi, 2017; Nizzi et al., 2012); as does the experience of *depersonalization*: a feeling of being "unreal" or "detached from oneself" (Yehuda et al., 2015). These beliefs and experiences can occur in the context of post-traumatic stress disorder, and when they do, the associated experience of dissociation or "shutting down" involves inhibition of emotion processing areas in the brain, including the amygdala (Yehuda et al., 2015). The "checking out" response can be contrasted with the (often coexisting, trading-off) response to trauma involving hyperarousal and emotional outbursts (Yehuda et al., 2015).

Most people -- estimates are around 50% to 70% of the population (Kessler, Sonnega, Bromet, Hughes, Nelson, 1995; Yehuda et al, 2015) -- have experienced a traumatic event, such as facing the threat of death, attack, molestation, rape, surviving or witnessing a horrible accident, experiencing combat. The great majority don't develop disabling PTSD (Yehuda et al, 2015), and purportedly don't experience a rupture in sense of self. Doris' theory makes important novel predictions relevant to traumatic experience. First, the more that a person's traumatic event involved profound cultural-level disturbances, or the person being prevented from expressing his or her values as a member of a group, the more he or she should report self-discontinuity -- as indexed by endorsement of "shutting down" experiences, associated trauma-related cognitions, and inhibited emotion: depersonalization, dissociation, and reports of a sense of a foreshortened future; and not hyperarousal. Furthermore, Doris' theory suggests that remediation of symptoms

will come about through a collaborative conversation about values, a position that proposes an intertwining of philosophy and clinical psychology -- and one that we support. Finally, it suggests unsettling effects on moral judgment of harm associated with traumatic experience. Specifically, harm of self and others may be judged as more acceptable to the extent that trauma causes "shut down" of emotional processing and an associated rupture in the sense of self -- as though one has "predeceased the body." This suggests a mechanism for inter- to intra-group spread of violence: targeted cultural annihilation may breed callousness broadly (not just retaliatory rage) because targeted, traumatized individuals experience affective shutdown that allows them to more easily harm close others, negatively affecting intra-group relations.

Future research consistent with Doris' pluralist account of moral agency has the potential to link thinkers across the disciplines of philosophy, psychology, and neuroscience. Ongoing investigations indicate that appropriate tools for these investigations will include measures that tap people's explicit endorsements of moral values (e.g. Clifford et al., 2015; Graham et al., 2011), sense of the self as continuous (self-discontinuity scale: Nizzi et al., 2012; Niemi & Nizzi, 2017), symptoms of avoidance and numbing, i.e., "shutting down" apart from hyperarousal after trauma (Clinician-Assisted PTSD Scale; Blake et al., 1995), and suicidality; as well as measures of neural activity and physiological markers of arousal (e.g., fMRI, EEG, EMG, EDA), implicit cognition, and emotional processing.

Doris' account acknowledges, in detail, deep constraints on human freedom. Happily, this theory of constraints has the potential to inspire much creative work, and to engender rich scientific and philosophic questioning about whether and how people exercise moral agency, and about the nature of the self.

References

Blake, D. D., Weathers, F. W., Nagy, L. M., Kaloupek, D. G., Gusman, F. D., Charney, D. S., Keane, T. M. (1995). The development of a clinician-administered PTSD scale. *Journal of Traumatic Stress, 8*, 75–90.

Clifford, S., Jewell, R. M., & Waggoner, P. D. (2015). Are samples drawn from Mechanical Turk valid for research on political ideology? *Research and Politics, Oct-Dec,* 1-9.

Crockett, M. J., Clark, L., Hauser, M. D., Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *PNAS, 107*(40), 17433–17438.

Doris, J. M. (2015). *Talking to Our Selves: Reflection, Ignorance, and Agency.* Oxford University Press: Oxford, UK.

Ehlers, A., & Clark, D. M. (2000). A cognitive model of posttraumatic stress disorder. *Behaviour Research and Therapy*, *38*, 319–345.

Foa, E. B., Ehlers, A., Clark, D. M., Tolin, D. F., & Orsillo, S. M. (1999). The posttraumatic cognitions inventory (PTCI): Development and validation. *Psychological Assessment, 11*(3), 303–314.

Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science*, *24*, 38-44.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*, 1029–1046.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*, 366-385.

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. (2007). The new synthesis in moral psychology. *Science, 316*(5827), 998-1002.

Kessler, R. C., Sonnega, A., Bromet, E., Hughes, M. & Nelson, C. B. (1995). Posttraumatic stress disorder in the National Comorbidity Survey. *Archives of General Psychiatry, 52*, 1048–1060.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature, 446*(7138), 908-911.

Miller, R. M., Hannikainen, I. A., & Cushman, F. A. (2014). Bad actions or bad outcomes? Differentiating affective contributions to the moral condemnation of harm. *Emotion, 14*(3), 573-587.

Niemi, L. & Young, L. (2016). When and why we see victims as responsible: The impact of ideology on attitudes toward victims. *Personality and Social Psychology Bulletin*, *42*(9), 227-1242.

Niemi, L. & Nizzi, M-C. (2017). Perceived self-discontinuity, PTSD and suicidality after sexual assault. *Manuscript in preparation.*

Nizzi, M-C., Demertzi, A., Gosseries, O., Bruno, M-A., Jouen, F., Laureys, S. (2012). From armchair to wheelchair: How patients with a locked-in syndrome integrate bodily changes in experienced identity. *Consciousness and Cognition, 21,* 431–437.

Park, G., Kappes, A., Rho, Y., Van Bavel, J. J., (2016). At the heart of morality lies neuro-visceral integration: Lower cardiac vagal tone predicts utilitarian moral judgment. *Social Cognitive and Affective Neuroscience*, *11*(10), 1588-1596.

Perkins, A. M., Leonard, A. M., Weaver, K., Dalton, J. A., Mehta, M. A., Kumari, V., Williams, S. C. R., Ettinger, U. (2013). A dose of ruthlessness: Interpersonal moral judgment is hardened by the anti-anxiety drug lorazepam. *Journal of Experimental Psychology: General, 142*(3), 612-620.

Yehuda, R., Hoge, C. W., McFarlane, A. C., Vermetten, E., Ruth A. Lanius, R. A., Nievergelt, C. M., Hobfoll, S. E., Koenen, K. C., Thomas C. Neylan, T. C., & Hyman, S. E. (2015). Post-traumatic stress disorder. *Nature Reviews, 1*, 1-21.