Manipulation, oppression, and the deep self

Forthcoming in *Behavioral and Brain Sciences*

Manuel R. Vargas

University of California San Diego, La Jolla CA, USA
mrvargas@ucsd.edu

Abstract: This essay considers various kinds of manipulation cases (local and global, dispositional and situational), and how Doris's Deep Self-style theory of responsibility fares in light of them. Agents acting with preferences adaptively formed under oppression are an especially interesting challenge for this sort of view, and the article considers what options may be available to Doris and others.

According to *Deep Self* theories, an agent is an apt candidate for blame when she acts in accord with values, self-governing policies, or particular higher-order desires (Frankfurt 1971; Watson 1975; Bratman 2007; Sripada 2015). Such theories are appealing because they allow us to distinguish between wayward or "alien" impulses and actions that reflect the agent's "true," "deep," or "real" convictions. They also capture the idea that one reason for blaming wrongdoers is that the wrongdoing expresses something about the wrongdoer. However, Deep Self views face an important objection: they deliver counterintuitive verdicts about moral responsibility in manipulation cases (McKenna & Pereboom 2016). To see why, consider an agent whose deep self is manipulated, unknowingly "implanted" with values or desires that replace her prior desires and values. Such agents can seem to be paradigmatically *not* responsible for actions derived from the manipulation. But on a Deep Self theory, the basis of responsibility just is whether the action reflects or expresses the values (or what have you) that the agent has. So it looks like the Deep Self theorist has to say that action that flows from the manipulated self is still responsible action. Manipulation cases—and, as I'll argue, some related but less science-fiction-y examples, including oppression and adaptive preferences—are a deep problem for the Deep Self approach.

Enter John Doris's wonderful recent book, *Talking to Our Selves* (2015). Doris offers an appealing upgrade to the traditional Deep Self account of responsible agency. He holds that responsible agency is present when an agent's actions are structured by values, or desires the agent accepts as determinative in practical planning. However, there is no requirement that the agent be aware of these values, or have self-consciously adopted them. What is required is only that the agents be willing to appeal to those desires or values in the justification of action plans. In keeping with Doris's emphasis on *collaborativism* (or the idea that optimal human rationality is socially embedded), he maintains that values can be discovered and even created in the social context of collaborative reasoning.

Can Doris's account overcome worries about manipulation? My suspicion is that the social dimension of Doris's account raises challenges that are particularly difficult to address within the confines of the Deep Self approach. Here, I consider several different kinds of manipulation (global and local, disposition- vs. situation-focused). A range of real-world cases suggest that it is difficult for his account to capture some familiar convictions about when oppression undermines culpability.

Manipulation cases are not a primary concern of Doris's book, but he does remark on a case of *global* manipulation, where an agent is subject to comprehensive and coercive value indoctrination. The model here is Patty Hearst's kidnapping and subsequent participation in the Symbionese Liberation Army in 1974. Doris argues, plausibly enough, that his account can handle such cases. Coercive indoctrination, he thinks, is likely to bring with it impairments in valuational capacities and value-expression, and even disruptions of personal identity (p. 31). Either Hearst's actions didn't express her values (if, for example, she was coerced or impaired), and thus she wasn't responsible, or her actions did express her values (without impairment), but then she was responsible for that reason.

What about cases of *local* value manipulation? Imagine a person who values being generous with comments on student research across all the usual contexts, but who is unknowingly subjected to a manipulation where that generosity is deleted—but only in workshop contexts and not in office hours, labs, or conferences. Is the post-manipulated person responsible for her insensitivity in the workshop context? I'd wager that many would say no. Or consider a case of a monogamous lover who, because of manipulation, now values infidelity, but only in a narrow context. Or, consider a once-relaxed commuter who, post-manipulation, is made particularly prone to road rage, but only on rainy days. Implanted or manipulated values seem like the wrong kind of basis for responsibility.

Akin to the global case, perhaps Doris will lean on the thought that upstream manipulations produce downstream impairments to valuational capacities (p. 32). It is an empirical question whether we will ever have the ability to narrowly manipulate an agent's values in the way I've suggested. It is also unclear why local manipulation would necessarily bring with it impairments to an agent's valuational capacities or personal identity, and if so, why those impairments would always be operative in just those contexts where the manipulated values are in play. At any rate, there is no obvious *conceptual* barrier to the possibility of a local value manipulation without impairment. To the extent to which we find local manipulation cases to be instances of non-responsibility, then the Deep Self theory gets the wrong verdict.

The cases described above have all been instances of *disposition-focused manipulation*, where what is manipulated is the agent's "in the head" psychological dispositions. What about "out of the head" manipulation cases, or *situation-focused* manipulations? Suppose I know you are desperate to feed your family and I offer you demeaning and potentially illegal work that I know you are only willing to take out of desperation and lack of other alternatives. Suppose, too, that we both know that the work will likely shift your values in a direction that better comports with that work. Suppose

further that I have some control over whether you have access to more palatable alternatives, and have conspired to ensure that you don't have access to those alternatives. Would you be fully responsible for the choice to take that job, and all that follows? Whatever the right answer is—and I suspect intuitions differ about choices under exploitatively engineered contexts—the fact that we can wonder about such cases seems puzzling, given a Deep Self view. Situation-focused manipulations look like they should be entirely irrelevant to moral responsibility, at least on a traditional Deep Self theory.

Oppressive social contexts may help bring out the stakes of the underlying puzzle. Let *oppression* be the property of unjust or immoral treatment, social relations, or distributions of opportunities, when it is produced by immoral or unjust social and political arrangements (Vargas forthcoming). Although oppression is not always an impediment to responsibility, it is sometimes part of the explanation for why it doesn't always seem to be the fault of desperate people when they do desperate things.

At first pass, there are plenty of things Doris might say on behalf of the Deep Self approach in contexts of oppression. Consider someone who reluctantly takes to, say, low-level drug trafficking because in his part of town the non-criminal ways of earning money are difficult to secure, involve considerable burdens (e.g., traveling through hostile neighborhoods and/or relying on lengthy and uncertain commutes), or come at particularly high social costs (risking estrangement and vulnerability to violence for not participating in peer-group activities). In such a case the reluctant dealer would not be acting in accord with his values, and to that extent would not be as culpable as he would be were he acting from his values. So here the Deep Self theory delivers the right verdict. Fair enough.

The challenge of situational manipulations is deeper, however. One way situational manipulations work is by modifying dispositions. Take the case of adaptive preferences. Adaptive preferences are preferences that are formed in response to restricted options (Elster 1983). The

particularly troubling cases of adaptive preferences are when the preferences are for things that are either counter to one's flourishing or otherwise not what one would prefer under more normatively optimal circumstances (Khader 2011). So, for example, in cases of domestic violence the abused partner may come to think of the abuse as merited or deserved. Or someone might come to think that because of her social identity (gender, race, social class, etc.), her labor deserves less compensation than it would were it done by someone with a different social identity.

Social orders inculcate norms that advantage some at the cost of others, and oppression plausibly relies on internalization for much of its efficacy. Value formation frequently occurs under conditions where people have an inadequate opportunity to deliberate upon and to choose morally palatable alternatives. If so, then the worry about adaptive preferences and the effects of oppression more generally on culpability are not readily addressed by simply consulting the offending agent's deep self. The worry just is that in the real world deep selves are too often the products of processes that are themselves culpability-undermining. Either we need a compelling error theory for these intuitions, or we need to give up the idea that responsibility is grounded in the history-insensitive valuational structure of an agent.

Doris could address these challenges by stipulating a historical condition on moral responsibility, as others have done (Fischer & Ravizza 1998; Mele 2009). However, this would be at odds with Doris's explicit strategy in the book, which eschews appeals to history in favor of appeals to an agent's occurrent psychological features (pp. 30-31). More importantly, he'd need some explanation of why such a requirement isn't an *ad hoc* departure from the basic explanatory strategy of the Deep Self theory. History might matter, but if the way it matters is antecedent to the presence or absence of the Deep Self, one might wonder whether the Deep Self is merely symptomatic of something else that actually grounds responsibility.

A different response could build on the collaborativist/socially-responsive element that animates Doris's particular approach. Perhaps Doris could maintain that in some sense, for all agents, it is adaptive preferences all the way down. If so, then there is nothing special about the apparently awkward cases; like all cases of responsibility, it is a matter of whether the putatively culpable action reflects the agent's deep self.

That's a principled reply, if a costly one. The spouse who thinks she deserves abuse and puts herself and her kids at risk would, according to such an account, be fully culpable because she takes her meriting abuse to be determinative in practical reasoning. The victim of wage and employment discrimination who fails to protest his treatment because he has internalized racial and class prejudice and thinks he doesn't deserve a well paying job is acting responsibly, says the theory, if he acts from internalized values. Perhaps it is an insight from philosophical theorizing that such agents enjoy no diminution of their responsibility. It would take a compelling story to overturn the widespread sense that oppression and adaptive preferences matter for responsibility.

It is unclear how to square our evident willingness to find some values and their formation as an inadequate basis for responsibility with the two chief features of Doris's account of responsibility, i.e., acknowledgement of the way situations shape dispositions and the idea that responsibility is grounded in a Deep Self. I'm somewhat more optimistic about accounts that ground responsibility in rational capacities (Vargas forthcoming). On such accounts, if the agent is insufficiently capable of recognizing and suitably responding to moral considerations, then wrongful actions (grounded in preferences adaptively formed under oppression) aren't instances of responsible agency for reasons of rational impairment. Mitigation or diminutions of responsibility are explained in terms of constraints on the ability of agents to recognize and respond to moral considerations. For Deep Self theorists,

however, to appeal to this sort of rational impairment is tantamount to abandoning the Deep Self approach.

Doris could appeal to his pluralism about responsible agency (pp. 12, 171-175), addressing such cases by appeal to resources that are not, as it were, "deep selfy." Such a strategy would suffer its own cost: if Doris has to appeal to non-Deep Self theories to shore up the Deep Self account, then it gets harder to insist that the Deep Self approach is a particularly helpful way to think about responsibility.

I'm not sure what the right answer is here, but I've no doubt Doris will find an insightful way forward.

References

Bratman, M. E. (2007). *Structures of Agency: Essays*. New York: Oxford University Press, USA.

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

Elster, J. (1983). *Sour Grapes: Studies in the Subversion of Rationality*, Cambridge: Cambridge: Cambridge University Press.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge University Press.

Frankfurt, H. (1971). Freedom of the Will and the Concept of a Person. *Journal of Philosophy, 68*(1), 5-20.

Khader, S. (2011). *Adaptive Preferences and Women's Empowerment*. Oxford: Oxford University Press.

McKenna, M., & Pereboom, D. (2016). *Free Will: A Contemporary Introduction*. New York: Routledge.

Mele, A. (2009). Moral Responsibility and History Revisited. *Ethical Theory and Moral Practice, 12*(5), 463-475.

Sripada, C. (2015). Self-Expression: A Deep Self Theory of Moral Responsibility. *Philosophical Studies, 175*(5), 1203-1232.

Vargas, M. (Forthcoming). The Social Constitution of Agency and Responsibility: Oppression, Politics, and Moral Ecology. In M. Oshana, K. Hutchinson, & C. Mackenzie (Eds.), *The Social Dimensions of Responsibility*. New York: Oxford University Press.

Watson, G. (1975). Free Agency. *Journal of Philosophy, 72*(8), 205-220.