Zooming irresponsibly down the slippery slope

Daniel Coren (Daniel.coren@colorado.edu)
Visiting Assistant Professor, McMaster University
Penultimate draft; final draft was accepted for publication in Analysis on December 2nd 2020
Please do not cite this draft.

Philosophers are familiar with slippery slopes and vague predicates. It would not surprise many to hear that, just as it’s not clear what number of single-hair-removals (if any) transforms a person from ‘not bald’ to ‘bald’, it’s also vague what number of additional seconds of age (if any) transforms someone from ‘not morally responsible’ to ‘morally responsible’. Nor would it surprise many to hear that, with mathematical induction, a plausible base case and induction step appear to show that there exists no age at which anyone is, or can be, morally responsible. But by using Boolos’s technique of generating spanning conditionals from universal instantiations of the induction premiss, I’ll suggest an underexplored diagnosis of some problems for the possibility of moral responsibility. If we zoom rather than merely slide down the slippery slope, we can see that, for instance, Galen Strawson’s Basic Argument and Susan Wolf’s Troubling Train of Thought reason in an unnatural direction: if \(a\) clearly has some property that results in our saying that \(a\) is \(F\), and if \(b\) less clearly has that property, then it is the case that \(b\) is \(F\).

Following Boolos (1991: 695), define mathematical induction as a principle that asserts that a predicate applies to every natural number if it applies to 1 and to the successor of every natural number to which it applies. Take ‘\(\prime\)’ to denote the successor function. Take ‘\(x\)’ to be a variable ranging over the natural numbers. Like Boolos, we’ll call the statement expressed by the result of inserting a predicate into the induction schema:

\[
\ldots 1 \ldots \land \forall x(\ldots x \ldots \rightarrow \ldots x' \ldots ) \rightarrow \forall x \ldots x \ldots
\]

induction with respect to that predicate. Let ‘\(M\)’ be the following predicate: ‘is a number of seconds of age (denoted by the variable ‘\(x\)’ ranging over the natural numbers) at which a
human being is or can be *morally responsible*. There are no one-second-old human beings who are morally responsible. So, we have our base case. As for the induction premise: The addition of one second to a human being’s age (and a corresponding slight increase in cognitive development, understanding and awareness) cannot transform that human being from not morally responsible to morally responsible. If a human being aged \( n \) seconds isn’t morally responsible then nor is a human being aged \( n+1 \) seconds. It seems true, then, that \( \forall x (\neg Mx \rightarrow \neg Mx') \). With our base case and our induction step we can conclude that nobody of any age is, or can be, morally responsible for anything.

There are three ways to proceed. First, one might accept the conclusion. That’s an unpopular option. Most of us seem to hold it to be true and deeply important that villains are blameworthy for their villainy and saints praiseworthy for their saintly deeds.\(^1\) Second, one might attack the base case. That’s an even less popular option: I do not know of anyone who holds the view that newborn infants are (even a little) morally responsible for anything. Third, one might attack the induction step. That is, you might argue that \( \neg \forall x (\neg Mx \rightarrow \neg Mx') \). This seems to be the most reasonable strategy. But it is difficult to show that \( \exists x (\neg Mx \rightarrow \neg Mx') \), that there is a specific number of seconds that is the last age at which a person may be held morally responsible for anything, such that, at the age denoted by that number’s successor, that person is suddenly a morally responsible agent.

---

\(^1\) There are, of course, sceptics and nihilists about moral responsibility. Neil Levy (2011) argues that we have no sound basis for praising or blaming anyone more than anyone else given the influence of luck in generating our characters and other causes of our decisions. Hard determinist reasoning, to which some such as Honderick (2002) and Pereboom (2001, 2014) are friendly, entails that we are not morally responsible.
But we might attack the induction premise by showing that among the logical entailments of its infinitely many universal instantiations are infinitely many ‘spanning’ conditionals, many of which are implausible. As Boolos (1991: 696-7) and Coren (2018: 98) point out: if an induction premise such as $\forall x (\sim Mx \rightarrow \sim Mx')$ is true, then so are all the members of the set $\{ (\sim M1[i] \rightarrow \sim M1[i+1]) : i > 1; i \rightarrow \infty \}$, which contains infinitely many universal instantiations of $\forall x (\sim Mx \rightarrow \sim Mx')$. Therefore, all spanning conditionals among those instantiations, including $(\sim M1 \rightarrow \sim M1[31])$, $(\sim M1 \rightarrow \sim M1[1,009,152,000])$ and $(\sim M1 \rightarrow \sim M1[2,018,304,000])$ are true. So, with those instantiations we get some reasonable results such as the following:

(Absolution from infancy to 31-seconds-older infancy): If a 1-second-old human being is not morally responsible for anything then a 31-second-old human being is not morally responsible for anything.

That seems sensible enough. But if we zoom rather than merely slide down the slippery slope, some of those infinitely many instantiations have extraordinary results:

(Absolution from infancy to adulthood): If a 1-second-old human being is not morally responsible for anything then a 1,009,152,000-second-old (a 32-year-old) human being is not morally responsible for anything.

(Absolution from infancy to middle age): If a 1-second-old human being is not morally responsible for anything then a 2,018,304,000-second-old (64-year-old) human being is not morally responsible for anything.

If at first the infant, then, ‘mewling and puking in the nurse’s arms’, is not morally responsible then nor is ‘the justice in fair round belly with good capon lin’d’ (Shakespeare’s words in *As You Like It*). Such conditionals seem false. Since they follow from the induction premise, we have good reasons to reject $\forall x (\sim Mx \rightarrow \sim Mx')$. Spanning conditionals such as (Absolution from infancy to middle age), unlike more modest conditionals such as (Absolution from infancy to 31-seconds-older infancy), show that $\forall x (\sim Mx \rightarrow \sim Mx')$ moves in the wrong direction. We normally reason: if $a$ clearly has some property that results
in our saying that \( a \) is \( F \), and if \( b \) even more clearly has that property, then it is the case that \( b \) is \( F \). We do not normally reason: if \( a \) clearly has some property that results in our saying that \( a \) is \( F \), and if \( b \) less clearly has that property, then it is the case that \( b \) is \( F \). Yet \( \forall x (\neg Mx \rightarrow \neg Mx') \) has us reason in the latter (unnatural) way. This unnatural reasoning in the induction step is not apparent when we consider its modest instantiations. But by applying Boolos’s elegant emphasis on some spanning conditionals, we get a better grip on what’s gone wrong.

I wish to apply these distinctions to some arguments that have enjoyed some attention in the literature on moral responsibility. Both arguments purport to refute the possibility of moral responsibility. The first is the so-called Troubling Train of Thought discussed with admirable care by Susan Wolf (2015):

\begin{align*}
\text{(T1) Either an action is out of character or in character.} \\
\text{(T2) If the action is out of character then we cannot be responsible for it (since we can only be responsible for actions that come from our character).} \\
\text{(T3) If the action is in character then we can only be responsible for it if we are responsible for the character from which those actions flowed (producing either an infinite regress or an acceptance of the fact that one’s ability to affect or change one’s character must ultimately flow from aspects of one’s character that one is not responsible for).} \\
\text{(T4) From T1-T3, it follows that we are not, and cannot be, morally responsible for our actions.}
\end{align*}

Wolf argues that T2 and T3 depend on faulty assumptions about the connections between character and responsibility. T3 assumes that character traits are like conditions such as hypnosis or addiction (2015: 357). One might also attack T2 by arguing that (moral) responsibility for action can obtain by satisfying control and knowledge conditions without satisfying a qualitatively deeper character condition. Another response is to analyse T3 as we analysed the induction step above. Consider some instantiations of T3: 1,009,152,000-second-old Jane responsible for her in-character actions at the 1,009,152,000\textsuperscript{th} second of her
life only if she’s responsible for the 1,009,151,999-second-old character from which those acts flowed.

... 1,009,151,997 one-step instantiations later…

2-second-old Jane is responsible for in-character actions at the 2nd second of her life only if she’s responsible for the 1-second-old character from which those acts flowed.

None of those one-step instantiations seems implausible. Though they reason in an unnatural direction, they do so in such small increments that the strangeness of this reasoning is not immediately apparent. They logically entail some reasonable spanning conditionals:

(Character from 40 to 35): 1,261,440,000-second-old (40-year-old) Jane is responsible for her in-character actions at the 1,261,440,000th second of her life only if she’s responsible for the 1,103,760,000-old (35-year-old) character from which those actions flowed.

But such seemingly innocuous one-step conditionals also logically entail this oddity:

(Character from adulthood to infancy): 1,009,152,000-second-old (32-year-old) Jane is responsible for her in-character actions at the 1,009,152,000th second of her life only if she’s responsible for the 1-second-old character from which those acts flowed.

(Character from adulthood to infancy) is deeply implausible. In particular, that spanning conditional is much less plausible than any of the 1,009,151,998 one-step conditionals it takes to descend from the 32-year-old’s character-responsibility to the newborn infant’s character-responsibility. Since T3 entails (Character from adulthood to infancy), we have a good reason to reject T3. We can generate plenty of other good reasons from the many other implausible spanning conditionals that spring from T3’s instantiations.

Like Wolf’s T1-T4, Galen Strawson’s (1994) Basic Argument purports to show that we cannot be (truly or ultimately) morally responsible:
(B1) The way an agent, Jane, acts at an age such as 32 is a function of how Jane is, mentally speaking, at 32.

(B2) In order to be truly morally responsible for her action at 32, Jane must be truly responsible for how Jane is, mentally speaking, at 32.

(B3) In order to be truly morally responsible for the way Jane is, mentally speaking, at 32, Jane must be truly morally responsible for earlier choices (such as choices made at 31) that brought about her state at age 32.

(B4) Heredity and previous experience, rather than any conscious, reasoned choice, determine the way we are, mentally speaking, at any age.

(B5) From B1-B4, we are not truly morally responsible for any of our actions.

Studying B3 in a way that’s analogous to the way we studied Wolf’s T3, we can construct a series of instantiations:

The way 1,009,152,000-second-old (32-year-old) Jane acts is a function of how Jane is, mentally speaking, as a 1,009,151,999-second-old.

…1,009,151,997 one-step instantiations later…

The way Jane is, mentally speaking, as a 2-second-old is a function of the way Jane was, mentally speaking, as a 1-second-old.

None of those one-step instantiations seems implausible. But, with a familiar technique, we can construct a more provocative spanning conditional:

(Blame the baby’s mental state): The way 1,009,152,000-second-old (32-year-old) Jane is, mentally speaking, is a function of the way Jane was, mentally speaking, as a 1-second-old.

Now, Strawson might wish to defend (Blame the baby’s mental state) and other such spanning conditionals. Whereas Wolf aims to make sense of some salient features of our ordinary conception of character and responsibility, Strawson argues that our ordinary notion of responsibility is confused. But my point is that Strawson’s B3 has instantiations

---

2 As Wolf (2015: 360) notes, Strawson does not discuss character. Instead, Strawson uses the more general phrase and accompanying concept of ‘the way we are, mentally speaking’.
that logically entail not only innocuous one-step conditionals but also some decidedly implausible spanning conditionals. Just as in the mathematical induction step in the initial argument, and just as with Wolf’s T3, Strawson’s B3 reasons unnaturally: if \(a\) clearly has some property that results in our saying that \(a\) is \(F\), and if \(b\) less clearly has that property, then it is the case that \(b\) is \(F\). This is easy to miss with modest instantiations. But zooming and (some) spanning conditionals help us to see what’s amiss.\(^3\)

### References


\(^3\) My analysis may count in favour of views such as Kane’s (1996) on which there are cases where Jane’s character-determining decisions are not determined and lack sufficient causes. Without a prior sufficient cause for which Jane must be responsible, there is no need for positing an induction step susceptible to a regress or strange spanning conditionals. Of course, Kane’s view is not uncontroversial: Pereboom (2001: 47-50), for instance, criticizes Kane’s view.


