Time-Slice Agency and Moral Responsibility

Jeff Sebo
New York University

Abstract

Many philosophers think that we must be one and the same agent across time in order to be responsible for what we did in the past. I argue that this is a mistake: even if we are not one and the same agent across time, we can still be morally responsible for what we did in the past in many ways, e.g. we can still be complicit in, criticizable for, and/or liable for what we did in the past. I then argue that this approach to personal responsibility, which analyzes personal responsibility on the model of interpersonal responsibility, explains and justifies our practice of holding each other responsible for different past actions in different ways at least as well as, if not much better than, more traditional approaches to personal responsibility do.

1. Introduction

Many philosophers think that agent identity is a necessary condition for moral responsibility. That is, they think that we must be one and the same agent across time in order to be responsible for what we did in the past.[1] In a narrow, technical sense of 'responsibility,' this is clearly right. If *x* did not break your window, then *x* is not directly causally, or morally, responsible for breaking your window. But our discourse and practice around moral responsibility involves much more than attributions of direct responsibility: It also involves attributions of complicity, indirect responsibility, criticizability, liability, and more. And when we take this broader discourse and practice into account, it is not at all clear that we must be one and the same agent across time in order to be morally responsible for what we did in the past. In fact, it might even be the case that our considered judgments about moral responsibility make more sense, rather than less sense, on the assumption that we are different agents across time.

My aim in this paper is to argue that a *fine-grained theory of agent identity*, according to which each agent persists for only a short period of time, explains and justifies our attributions of moral responsibility at least as well as, if not much better than, a *course-grained theory of agent identity*, according to which each agent persists for a long period of time, does. I will proceed as follows. In section 2, I will consider several aspects of our discourse and practice around moral responsibility and show how, in each of these respects, we can be morally responsible for an action without having performed it. In section 3, I will argue that, even if we accept a *time-slice*

[1] See, for example, DeGrazia (2005), Glannon (1998), Maddell (1981), Olson (1997), and Schechtman (1996). (Note that not all of these philosophers think that agent identity has the logic of numerical identity.) And for a critique of this view about agent identity and moral responsibility that complements my argument here, see Shoemaker (2012).

theory of agent identity, according to which each agent persists for only a few seconds, these considerations are more than enough to make us morally responsible for most, if not all, of our past actions. Finally, in section 4, I will show how this approach to personal responsibility works in practice, by considering a series of cases and arguing that a time-slice theory of agent identity explains and justifies our judgments about these cases at least as well as, if not much better than, other, more traditional theories of agent identity do.

Before I begin, a couple clarifications about my strategy in what follows. First, I will use 'agent' exclusively in the moral sense of the term and 'person' exclusively in the metaphysical sense of the term, so as to make it clear that this paper is about morality rather than metaphysics. On this terminology, then, it will be conceptually possible for agents and persons to come apart. Indeed, on this terminology, the idea that agents and persons come apart is the same as the idea that personal identity is not "what matters in morality" (Parfit 1984, pp. 321-50). Second, even though this paper is not about personal identity, it will still be useful for me to be able to talk about persons. So, I will assume that a standard theory of personal identity is true. That is, I will assume that physical, psychological, and narrative continuity are jointly sufficient for personal identity, and therefore I will assume that persons can persist for many years.[2] But nothing in my moral argument here will depend on this assumption. So, if you think that personal identity is more fine-grained than this, then you can translate all of my claims about persons into claims about human beings without affecting the substance of my argument at all.

[2] See Olson (1997) for a defense of a physical continuity theory of personal identity, see Parfit (1984) for a defense of a psychological continuity theory of personal identity and what matters in morality, and see Schechtman (1996) for a defense of a narrative continuity theory of personal identity and what matters in morality.

2. Our Discourse and Practice Around Moral Responsibility

I begin by reviewing several aspects of our discourse and practice around moral responsibility and showing how, in each of these respects, an agent can be morally responsible for an action without having performed it.

2.1. Collective Responsibility

First, many philosophers believe that, when a group of agents think and act together, they constitute themselves as parts of the same *collective agent*, and therefore they count as *collectively responsible* for what they do together as well as *individually complicit* in what they do together.

What does it take to be part of the same collective agent? There are many answers to this question in the literature. For purposes of my discussion here, I will focus on a standard theory of collective agency, the *shared intention* theory, as developed by Michael Bratman (1999).[3] Here is how this theory works. Suppose that a group of agents build a house together. Do they count as a collective agent? That depends on why they do what they do. If each is acting completely independently of the others, and all of this activity just happens to result in a house being built, then they do not count as a collective agent. But if (as is much more likely) each agent does what she does at least in part because she *intends* for them build a house together, then they count as a collective agent, at least with respect to this particular activity.

[3] For alternative conceptions of shared agency (all of which are compatible with my claims here), see Gilbert (1990), Kutz (2000), and Velleman (1997).

It is important to note how flexible collective agency is. For example, a group of agents can share an intention to build a house even if they each act at different *times*, even if they each act in different *ways*, even if they each have different *reasons* for sharing this intention, and even if they are not consciously aware of sharing this intention at all. For example, a group of agents can share an intention to build a house even if one of them finances the project, another buys the materials, another builds the foundation, another puts up the walls, and so on. Further, they can share this intention even if one of them is doing it for the money, another is doing it for the exercise, and so on. Finally, they can share this intention even if they are not consciously aware of sharing it at all: as long as the best explanation of their behavior is that each of them intends, among other things, that they build a house together, they count as sharing an intention to build a house together, whether they know it or not.

Now, insofar as a group of agents constitutes themselves as a collective agent, we can draw at least two kinds of conclusion about moral responsibility. First, we can say that the group, as a collective agent, is *collectively morally responsible* for what it does. Thus, for example, we can say that a team of assassins is collectively morally responsible for killing their target, and we can therefore permissibly praise or blame the group as a whole for this action. Second, we can say that each member of the group is *individually morally complicit* in what the group does. What does the idea of individual complicitity amount to? The basic idea, as developed by Christopher Kutz (2000), is that each member of a group is individually morally responsible for her *participation* in what the group does. Thus, for example, we can say that each member of the team of assassins is individually morally responsible for her participation in the assassination, even if they contributed at different times, in different ways, and so on; and we can therefore permissibly praise or blame each member of the team for her participation in this action.

Of course, there are important questions about what exactly individual responsibility for participation in a collective action amounts to, in practice. Are you praise- or blameworthy for the entire collective action, such that everyone is responsible for the planning, the shooting, the getaway, and so on? Are you praise- or blameworthy only for your *contribution* to the collective action, such that the planner is responsible only for planning, the shooter is responsible only for shooting, the getaway driver is responsible only for getaway, and so on? Or, as seems plausible in practice but difficult to spell out in theory, are you praise- or blameworthy for "less" than the entire collective action but for "more" than your contribution to it? I will not try to answer this question here. Instead, I will simply assume that, when we think and act collectively, we are collectively responsible for what we do as well as individually complicit in what we do, where individual complicity involves, *at the very least*, individual responsibility for our contribution to our collective action.

2.2. Indirect Responsibility

Second, many philosophers believe that we can be praise- or blameworthy for the *foreseeable consequences* of our actions.[4] Thus, if a foreseeable consequence of one of your actions is that *someone else* acts wrongly, then you can be blameworthy for foreseeably causing or allowing them to act this way.

For instance, suppose that you hand me a gun, or even let me pick up a gun, with the foreseeable consequence that I will shoot somebody. Now suppose that, as a matter of fact, I

[4] For discussion of indirect responsibility in morality and the law, see Feinberg (1984) and Hart and Honore (1985).

shoot somebody. In this case, many philosophers would say that you are blameworthy for your action. This does not mean, of course, that you are blameworthy for *shooting someone* (I am blameworthy for doing that), nor does it mean that you are just as blameworthy as I am (foreseeably causing or allowing me to shoot someone does not warrant as much blame as shooting someone does, at least not in standard cases). Nevertheless, your action in this case warrants at least some blame. This is especially true if you cause or allow me to shoot someone intentionally, i.e., if you hand me the gun or let me pick it up, fully expecting me to shoot somebody as a result. But it is also true if you do it negligently, i.e., if you hand me the gun or let me pick it up never suspecting that I might shoot somebody as a result, even though you clearly should have.

2.3. Criticizability

Third, many philosophers believe that we can be *criticizable* for our character, i.e., for who we are and what we would do if given the chance. How is this different than blameworthiness? It involves similar reactive attitudes, but without the implication that we have done anything wrong.[5]

Consider, for example, a case of moral luck.[6] You and I are identical twins. We have the same beliefs, desires, aims, habits, and so on, and, as a result, we are likely to do the same things

---

[5] As Christopher Kutz (2000) puts the point:

> The central distinguishing feature of [accountability warranted by] reasons of character is that the relation between agent and harm need not be mediated by either causality or intentional conduct. Harms may be symbolic, standing for elements of character in agents other than those who brought them about (p. 43).

[6] For more on moral luck, see Williams (1981) and Nagel (1979).

in the same situations. One night, we each choose, completely independently of each other, to drive drunk. You get home safely, but I run over a young child. We can grant that you are not directly or indirectly responsible for my running over a child. Nevertheless, you would have done the same thing in the same situation; it was only a matter of luck that the child appeared in front of my car and not in front of yours. Thus, intuitively, there is a sense in which you and I are equally *criticizable*, and there is also a sense in which my action *reveals* this criticizability to people who know how similar we are. For example, suppose that our family members know that we have the same dispositions in general, and, as a result, they know that my actions provide evidence of your dispositions, and vice versa. In this case, our family members might be justified in criticizing you in light of what I do. For example, they might be justified in saying to you, "What does it say about you that your *identical twin* would drink and drive and kill a child?" If they said this, they would not be blaming you for my action. Rather, they would be criticizing you for your disposition to do the same thing in the same situation – as is evidenced by my doing so, and our having the same dispositions in general.

Moreover, this point holds to a degree even if you and I are different in some respects, and these differences are enough to make it the case that I would decide, all things considered, to act wrongly, whereas you would decide, all things considered, not to. For example, suppose that there are at least a few situations where I would choose to drink and drive and you would not, though you would still be very tempted to. In these situations, intuitively, I am a bit more criticizable than you are, though you are still criticizable to a degree, and my actions still *reveal* your criticizability to a degree. For example, if I choose to drink and drive in a situation where you would not have, our family members might still be justified in saying to you, "Sure, you might not have made the *same exact choice* in the *same exact situation*, but you still would have

been tempted to. And furthermore, if the situation was even a little bit different, you might have caved. So, this is a wake-up call for all of us. You need to work harder to reinforce your disposition not to drink and drive, and we need to work harder to support you in this – and maybe be a bit more wary around you in the meantime." In this case, your family members would not be blaming you for my action, nor would they even be criticizing you for your disposition to do the same thing in the same situation. Rather, they would be criticizing you for your disposition to do the same thing in *different* situations, as well as for your disposition to at least be *tempted* to do the same thing in the same situation – as is evidenced by my doing so, and our having many of the same dispositions in general.

2.4. Liability

Fourth, many philosophers believe that we can be *liable* for what other agents do. This means that we can have an obligation to *take* responsibility for what they do, and others can be justified in *holding* us responsible for what they do (or otherwise harming or benefiting us as a result of what they do), even if we did not perform the relevant action, even if we did not foreseeably cause or allow the relevant action, and even if we would not have done the same thing in the same situation (and would not even have been tempted to). The concept 'liability' has a legal as well as a moral sense. For present purposes, I will restrict my focus to the moral sense. Specifically, I will discuss three ways in which we can be morally liable for what others do.

2.4.1. Commitments

First, we can *commit* to being liable for what another agent does. For example, you might agree to serve as representative of a group that you are a member of. When you play this role, you agree to take responsibility, on behalf of the group, for what other members of the group say or do. For example, if you agree to become President of the United States, then this involves making a commitment to take responsibility, on behalf of your country, for what other Presidents have said and done on behalf of your country. Similarly, if you agree to become a parent, then this involves making a commitment to take responsibility, on behalf of your children, for what they say or do in general. Thus, for example, the President might find that she has a moral obligation to apologize, on behalf of her country, for a war that she was not directly or indirectly responsible for prosecuting (and would not have even been tempted to prosecute, if in the same situation). Similarly, a parent might find that she has a moral obligation to apologize, on behalf of her children, for a broken window that she was not directly or indirectly responsible for throwing a rock through (and would not have even been tempted to throw a rock through, if in the same situation).

2.4.2. Fairness

Second, you can have a *fairness-based* obligation to take responsibility for what other agents do, whether or not you are directly or indirectly responsible for the relevant actions, would have done the same thing in the same situation (or even been tempted to), or made a commitment to take responsibility for it. For example, Henry Shue (1999) argues that people currently living in

industrialized nations benefit from the effects of industrialization, and therefore they have a duty to compensate others who were, or will be, harmed by the effects of industrialization. Or, to take a simpler example: suppose that you inherit a large sum of money, and then you discover, after the fact, that your parents earned this money in the drug trade. Many people would claim that you have a duty to return the money or, if this is impossible, to donate it to a charitable cause, ideally one that benefits victims or survivors of the drug trade.

2.4.3. Punishing or Harming the Innocent As a Byproduct of Punishing the Guilty

Finally (this is not a sense in which we use the terms 'liability' or 'responsibility,' but it falls under the general heading of liability as defined here), we are sometimes justified in punishing or harming some agents as a *byproduct* of punishing others, whether or not the former agents are directly or indirectly responsible for the relevant action, would have done the same thing in the same situation (or even been tempted to), made a commitment to take responsibility for it, or benefited at all from it. This is true in an epistemic as well as in a practical sense.

First, we are sometimes justified in punishing the innocent as a byproduct of punishing the guilty on epistemic grounds. In particular, we often have to decide whether to punish somebody for a crime without knowing for sure if they committed it. This of course a difficult decision, because if we reserve punishment for people we *know for sure* are guilty, then we will inevitably let many guilty people go free; yet if we allow ourselves to punish people for crimes *without* knowing for sure that they committed them, then we will inevitably punish at least some innocent people. So the question is: how many guilty people are we willing to let go free for each innocent person we punish? And the answer is usually: a lot, but not an infinite amount.

(After all, we assume, we need to punish at least *some* guilty people.) Consequently, we settle on a policy for punishment whose foreseeable consequence is that we will accidentally punish at least *some* innocent people. Of course, it is tragic whenever this happens. But that does not mean that we are wrong to settle on this policy, or to punish people as the policy dictates. Rather, it just means that we live in a complicated world, we have to punish at least some guilty people, and we often have to make decisions about punishment without full information about the case.

Second, we might have to harm the innocent as a byproduct of punishing the guilty for *practical* reasons. For example, suppose that a man is a good father but a bad husband, and his wife has to decide whether to leave him, thereby punishing her husband *as well as* harming her children, or stay with him, thereby sparing her children *as well as* her husband. Either way, as before, this is a tragic choice. But it is also a choice that many of us have to make, since, again, we live in a complicated world, and we will not always be able to make decisions about blame and punishment in a social vacuum.

To be clear: I am not claiming here that, if we are justified in punishing or harming one agent as a byproduct of punishing another, then we would, or should, publicly proclaim that the former agent is *responsible* for anything. (At the very least, this would be a surprising extension of our current use of 'responsibility' in the interpersonal case, even though it does fall under the general heading of liability as defined here.) Instead, my only claim here is that we are sometimes justified in punishing or harming one agent as a byproduct of punishing another, for epistemic as well as practical reasons. And, as we will see below, this observation will be useful for purposes of explaining and justifying many of our considered judgments about responsibility in the intrapersonal case (whether or not we think, on reflection, that this observation is an observation about 'responsibility' in the traditional sense at all).

3. Personal Responsibility

I will now argue that, even if we accept a time-slice theory of agent identity according to which each person becomes a new agent every few seconds, these considerations are more than enough to make us morally responsible for most, if not all, of our past actions as persons. The key, I will argue, is to draw on the familiar analogy between nations and persons in order to show how the pressures of intrapersonal living, like the pressures of intranational living, conspire to make us think and act in a way that results in our being complicit in, criticizable in light of, and/or liable for most, if not all, of our past actions as persons.[7]

Consider the intranational case first. Specifically, consider the situation that American Presidential administrations are in, given that (a) they govern the nation sequentially and (b) each administration has only 4-8 years to do everything that they want to do. In this kind of situation, if the current administration wants to accomplish anything at all, then they have no choice but to think in terms of what came before and what will come after. What would be continuous with what at least some past administrations have done? And what would be continuous with what at least some future administrations are likely to do? These questions are important to ask because nations evolve gradually, and, even though they might be likely to evolve in a particular direction in the long run (say, towards more progressive policies), they are also likely to experience many small social and political fluctuations along the way (say, between somewhat more progressive Democratic administrations and somewhat more conservative Republican administrations). So each administration has a delicate balance to try to strike: on one hand, they have to try to enact policies that will push the nation in the direction that they want it to go in. On

[7] For two very different treatments of this analogy, see Plato (1997) and Parfit (1986), pp. 211-3.

the other hand, they have to try to enact policies that future administrations are likely to accept and build on – or at least not able to easily undo. What this means in practice is that each administration has no choice but to accept many of the policies put in place by past administrations and focus on making a relatively small number of changes – an approach which necessarily makes each administration complicit in what many past administrations have done as well as in what many future administrations are likely to do. And obviously, this is to say nothing of the sheer force of social and political inertia, which, in practice, will also provide very strong incentive for each administration to continue the policies started by past administrations.

Of course, this complicity will vary from case to case. In some cases the current administration might be complicit with past and future administrations *across* party lines. For example, insofar as Democratic and Republican administrations have the same values and/or would rather compromise than compete, they will all pursue the same temporally extended agenda. In other cases the current administration might be complicit with past and future administrations only *within* party lines. For example, insofar as Democratic and Republican administrations have different values and would rather compete than compromise, Democratic administrations will pursue one temporally extended agenda and Republican administrations will pursue another, conflicting temporally extended agenda. And so on. And in this kind of way, even though each administration is, of course, directly praise- and blameworthy only for what it does, we can nevertheless justifiably hold them responsible for a variety of past and future executive actions in a variety of ways. For example, we can praise and blame America for what it does, and we can praise and blame particular administrations for their complicity in and indirect responsibility for what America does (when appropriate) as well as hold them liable, at least in some ways, for what America does. We can praise or blame the Democratic and

Republican Parties for what they do, and we can praise and blame particular Democratic and Republican administrations for their complicity in and indirect responsibility for what their parties do (when appropriate) as well as hold them liable, at least in some ways, for what their parties do. And then we can also, of course, criticize individual Presidents in light of what past and future Presidents do when appropriate.

Now then: Assuming for the sake of argument that we accept a time-slice theory of agent identity, the intrapersonal case has the same kind of structure as the intranational case. Specifically, (a) our temporal selves govern the person as a whole sequentially, and (b) each temporal self has only, say, 4-8 seconds to do everything that it wants to do. (The precise duration of temporal selves will not matter for our purposes here.) In this kind of situation, if your current self wants to accomplish anything at all, then they have no choice but to think in terms of what came before and what will come after. What would be continuous with what at least some past selves have done? And what would be continuous with what at least some future selves are likely to do? As in the intranational case, these questions are important to ask because persons evolve gradually, and, even though they might be likely to evolve in a particular direction in the long run (say, towards a particular balance between work, family, friends, and so on), they are likely to experience many small psychological fluctuations along the way (say, between a "work self" who cares a bit more about work, a "family self" who cares a bit more about family, a "friend self" who cares a bit more about friends, and so on). So each temporal self has a delicate balance to try to strike: on one hand, they have to try to enact policies that will push the person in the direction that they want it to go in. On the other hand, they have to try to enact policies that future temporal selves are likely to accept and build on – or at least not able to easily undo. What this means in practice is that each temporal self has no choice but to accept

many of the policies put in place by past selves and focus on making a relatively small number of changes – an approach which necessarily makes each temporal self complicit in what many past selves have done as well as in what many future selves are likely to do. And obviously, this is to say nothing of the sheer force of social and psychological inertia, which, in practice, will also provide very strong motivation for each temporal self to continue the policies started by past selves.

Of course, as in the intrapersonal case, this intrapersonal complicity will vary from case to case. In some cases your current self might be complicit with past and future selves *across* social contexts. For example, insofar as your work self, family self, friend self, and so on have the same values and/or would rather compromise than compete (which is true *much* more often than not), they will all pursue the same temporally extended agenda. In other cases your current self might be complicit with many past and future selves only *within* the present social context. For example, insofar as your work self, family self, friend self, and so on have different values and would rather compete than compromise, your temporal selves at work will pursue one temporally extended agenda, your temporal selves at home will pursue another, conflicting temporally extended agenda, your temporal selves at the bar will pursue another, conflicting temporally extended agenda, and so on. And in this kind of way, even though each current self is, per our assumption here, directly praise- and blameworthy only for what it does, we can nevertheless justifiably hold it responsible for a variety of your past and future actions in a variety of ways. For example, we can praise and blame you as a person for what you do, and we can praise and blame your current self for its complicity in and indirect responsibility for what you do (when appropriate) as well as hold it liable, at least in some ways, for what you do. We can praise or blame your work self, family self, friend self, and so on for what they do, and we

can praise and blame your current self for its complicity in and indirect responsibility for what these selves do (when appropriate) as well as hold your current self liable, at least in some ways, for what these selves do. And then we can also, of course, criticize your current self in light of what your past and future selves do when appropriate.

Moreover, I should emphasize that the patterns of complicity, indirect responsibility, criticizability, and liability that emerge from this kind of dynamic are likely to be *much stronger* in the intrapersonal case than in the intranational case. After all, in the typical intrapersonal case, our temporal selves are highly physically, psychologically, and narratively continuous with each other. Thus, they are likely to be *complicit* in most of what each other does, since they constitute themselves as a temporally extended collective agent simply by acting on the beliefs, desires, aims, and so on that they share. They are likely to be *indirectly responsible* for most of what each other does, since they naturally first-personally anticipate what each other will do. They are likely to be *criticizable* in light of most of what each other does, since they have many of the same dispositions. Finally, they are likely to be *liable* for most of what each other does since (a) they have to commit to taking responsibility for most of what each other does in order to engage in most of the projects and relationships that make life worthwhile, (b) they are likely to benefit as a result of what each other does, and (c) we are likely to be justified in punishing or harming some of them as a byproduct of punishing others, since it is very hard for us to tell which temporal selves are complicit in which temporally extended actions, and it is also very hard for us to punish some temporal selves without also, thereby, harming others. And in all of these ways and more, the pressures of intrapersonal living are likely to generate a very strong presumption in favor of "full responsibility" for many past actions (i.e., blameworthiness, criticizability, and liability).

This approach to personal responsibility has powerful theoretical implications: not only does it explain and justify the plausible and widely accepted idea that we are morally responsible for most, if not all, of our past actions, but it *also* explains and justifies the plausible and widely accepted idea (among philosophers, at any rate) that personal identity, defined in terms of physical, psychological, and/or narrative continuity, plays a central role in explaining this fact. To see why, consider each of these relations in turn. First, if your present self is *physically continuous* with your past and future selves, then they will have to compromise and coordinate very often, and with respect to very many issues, if they want to accomplish anything at all. Second, if your present self is *psychologically continuous* with your past and future selves, then they will have many of the same thoughts, feelings, and habits in general, and they will also be able to compromise and coordinate easily and effectively, simply by first-personally remembering and anticipating what each other thinks, feels, and does. Finally, if your present self is *narratively continuous* with your past and future selves, then they will all think and act "as one" by default, simply by thinking and acting from the standpoint of the protagonist of their shared self-narrative. Thus, on a time-slice theory of agent identity, we should expect that personal identity, defined in terms of physical, psychological, and/or narrative continuity, would affect our moral responsibility for our past actions. However, we should not necessarily conclude, as many philosophers do, that these metaphysical relations *ground* our moral responsibility for our past actions. Instead, we should say that *moral* relations like complicity, criticizability, and liability ground our moral responsibility for past actions, and that metaphysical relations like physical, psychological, and narrative continuity affect our moral responsibility for past actions to the degree that they affect these moral relations.

Of course, even if a time-slice theory of agency implies that we are morally responsible for most, if not all, of our past actions in at least *some* of the senses discussed above, it will not necessarily imply that we are morally responsible for most, if not all, of our past actions in *all* of the senses discussed above. After all, just as a particular maverick Presidential could conceivably break free from all of the physical, psychological, and narrative pressures towards conformity and chart a radically new course for the nation as a whole, your current self could conceivably break free from all of the physical, psychological, and narrative pressures towards conformity and chart a radically new course for the person as a whole. And in this kind of case, a time-slice theorist would claim that, while your current self might be *criticizable* in light of many of your past actions (at least to a degree) as well as *liable* for most, if not all, of your past actions (at least to a degree), they are not *praise- or blameworthy for* or even *complicit in* any of your past actions. This might seem like a radical view. But as I will now argue, many of us already accept this view in everyday life, as is demonstrated by the subtlety and variety of our attributions of moral responsibility for past actions.

4. What Happens in Vegas Stays in Vegas (Kind of)

I will now show how this approach to personal responsibility works in practice, by considering a series of cases and arguing that a time-slice theory of agent identity (coupled with standard assumptions about moral responsibility) explains and justifies our judgments about these cases at least as well as, if not much better than, other, more traditional theories of agent identity (coupled with standard assumptions about moral responsibility) do.

Consider first a typical case of infidelity: A man loves his family, and he thinks of himself as fully committed to them. But he also feels a bit stifled by them, and, as a result, his commitment to them is a bit weaker than he thinks – not so weak that he consciously plans to have an affair, but weak enough so that he *unconsciously* plans to have one, i.e., he does everything that he does at least in part because it might lead to an affair. For example, he eats well, dresses well, goes to the gym, and so on at least in part so that he can attract potential partners; he goes to the bar with his friends after work at least in part because he might meet someone there, and so on. Then, one night, it all pays off: he meets someone, they start talking, one thing leads to another, and he has an affair. Then he destroys the evidence, goes home, and lies to his family about where he was and what he did. Finally, he settles back into his normal routine. He never has another affair, but this is just by chance. He often thinks fondly about the affair that he had, and he fully intends to have another one if given the chance (though, again, he might not consciously realize this).

In this case, a time-slice theory of agent identity implies, along with other theories, that the man is fully morally responsible for the affair (i.e., blameworthy, criticizable, and liable for the affair) even when he comes back home. After all, all of his temporal selves share an intention to have an affair (even if they contribute to it at different times, in different ways, with different levels of awareness, and so on) and, therefore, they are collectively responsible for it and individually complicit in it. Moreover, they are all complicit in activities that foreseeably caused/allowed the affair to happen, and therefore they are collectively *indirectly* responsible for it as well. (They might not have foreseen the affair, but they should have.) Further, they all have many of the same character traits and *would* have done the same things in the same situations; and therefore they are all *criticizable* in light of the affair. Finally, they are all *liable* for the

affair, in all the ways we considered: they are all committed to taking responsibility for what each other does, they all benefited from the affair (by participating in it and/or by first-personally remembering it), they are all accountable for the affair – in part because other people might not be sure which temporal selves are guilty, and in part because it would be hard for other people to punish some temporal selves without also, thereby, harming others. Thus, a time-slice theory of agent identity implies that, in this case (and others like it), the man cannot permissibly defend himself, when he gets home from having an affair, by saying that "someone else" did it and leaving it at that. It may be true that other temporal selves slept with someone else, but his current self is still blameworthy, criticizable, and liable for this behavior in many ways.

Now consider a second case of infidelity. A man loves his family, and he is *usually* fully committed to them. But when he goes to the bar with his friends, a different side of him comes out. In this context, he still loves his family, and he still *thinks* of himself as fully committed to them, but he also feels a bit stifled by them, and, as a result, his commitment is a bit weaker than normal – not so weak that he consciously plans to have an affair, but weak enough so that he *unconsciously* plans to have one, i.e., he does everything that he does, at the bar, at least in part because it might lead to an affair (though he may not realize this). Most of the time, of course, nothing happens: he goes home, regains his normal perspective, and feels grateful to be back with his family again. But then, one night, he goes to the bar and meets someone new, they start talking, one thing leads to another, and he has an affair. Then he "comes to his senses," feels terrible about what happened, destroys the evidence, and goes home and lies to his wife – not because he wants to keep having this or any other affair, but because he loves his family and worries about losing them if they find out. Finally, he settles back to his normal routine. He never has another affair, but this is just by chance. As before, he is *usually* fully committed to his

family, but he still, on occasion, places himself in contexts where this other side of his character comes out – and in these moments, if he met the right person, then he would have another affair.

This case is similar to the first, but we have to treat it a bit differently. In this case, we have to distinguish the temporally extended collective agent who had an affair from the temporally extended collective agent who did not. The former, temporally extended "cheating self" is responsible for the affair in all the ways that I just mentioned: they are blameworthy, criticizable, and liable for it. In contrast, the latter, temporally extended "normal self" did not share an intention to have an affair, and therefore they are not complicit in it. However, they did share an intention to *cover up* the affair, and therefore they are complicit in the cover up; and they are also *indirectly* responsible for the affair as well as the cover up. Moreover, they have many of the same character traits as the cheating self, and would have been at least *tempted* to do the same thing in the same situation, and therefore they are at least *somewhat* criticizable in light of the affair. Finally, they are all liable for the affair in all of the ways that I just mentioned, except for one: unlike before, they feel terrible about the affair and, as a result, benefit less from the memory of it. All told, then, a time-slice theory of agent identity implies that, in this case as well (and others like it), the man cannot permissibly defend himself, when he gets home from having an affair, by saying that someone else did it and leaving at that. It may be true that other temporal selves had an affair, but his current self is still blameworthy for the cover up, criticizable in light of the affair (to a degree), and liable for the affair (to a degree).

Now consider a third case of infidelity. A man loves his family, and he is always fully committed to them. But then, one day, he goes on a trip to Vegas (something that he never does) and is surprised to find that this trip brings out a new side of his character. Specifically, while on the trip, he still loves his family, and he still *thinks* of himself as fully committed to them, but he

also feels a bit stifled by them, and, as a result, his commitment is a bit weaker than normal – not so weak that he consciously plans to have an affair, but weak enough so that he *unconsciously* plans to have one, i.e., he does everything that he does in Vegas at least in part because it might lead to an affair. And, sure enough, he meets someone during the trip, they start talking, one thing leads to another, and he has an affair. Then he "comes to his senses," realizes what he did, and resolves to do the right thing: he goes home, tells his wife what happened, apologizes profusely, and promises to do everything that he can to make it up to her, as well as to make sure he never places himself in another situation where this other side of him might come out and have another affair. And this is exactly what he does.

This case is similar to the previous one, but we have to treat it a bit differently as well. As in the previous case, we have to distinguish the temporally extended collective agent who had an affair from the temporally extended collective agent who did not. And, as in the last case, we can say that the former, temporally extended "cheating self" is responsible for the affair in all the ways that we have discussed: they are blameworthy, criticizable, and liable for it. However, we should say that the latter, temporally extended "normal self" is morally responsible in only some respects. They did not share an intention to have an affair or to cover it up, and therefore they are not complicit in either of these things. Moreover, they could not have foreseen that the trip to Vegas would lead to an affair (since, unlike in the first two cases, the man was not even aware that this side to his character existed), and therefore they are not *indirectly* responsible for the affair either. However, they do share many character traits with the cheating self and would have at least been *tempted* to do the same thing in the same situation, and therefore they are still at least somewhat *criticizable* in light of the affair. Moreover, they are still somewhat *liable* for the affair, in the same respects as in the second case: they have an obligation to take responsibility,

on behalf of the person as a whole, for the affair, and other people are justified in *holding* them responsible for the affair, for epistemic as well as practical reasons. Thus, even in this latter, more exceptional case, a time-slice theory of agent identity implies that the man cannot permissibly defend himself, when he gets home from having an affair, by saying that someone else did it and leaving at that. It may be true that other temporal selves had an affair, but his current self is still criticizable and liable for it (at least to a degree), and this will have to take priority when he talks to his family about the affair.

I believe that all of these results are plausible. Intuitively, there really is a sense in which this man's current self is more responsible for the affair in the first case than in the second, and in the second case than in the third. Yes, in all three cases the man should apologize profusely for the affair and promise to do everything he can to make up for it and never let it happen again. But provided that he does this, it seems completely reasonable for him to add, in the second and third cases, that "I was someone else when I did this, and the *real me* would never do this to you" where this statement marks a difference between the "cheating self" who had the affair and the "normal self" who is taking responsibility for it now, and also makes it clear that his current, normal self would not have done the same thing in the same situation. Likewise, it seems completely reasonable for him to add, in the third case, that "I never even knew that side of me *existed*," where this statement makes it clear that his current, normal self could not have foreseen that his cheating self would have this affair, though he can now.

Indeed, this kind of talk about responsibility is common. People often say "I was someone else when I did this," "I was not myself when I did this," and so on, in order to draw a line between the side of them who performed the relevant action and the side of them taking responsibility for it now – not in order to escape all responsibility, but rather in order to clarify

the sense in which they *are* responsible. And of course, at least part of the reason why people make these claims is that people often respond well to them. For instance, we can easily imagine a person saying, after a moment of critical reflection, "I hate the side of you that cheated on me, but I still love the real you, and I know that the real you would never hurt me like this." Granted, we can also easily imagine her adding: "Still, I think that we should get a divorce," since, after all, her choice is between leaving her husband, thereby sparing his normal self as well as his cheating self, and leaving him, thereby punishing his cheating self as well as harming his normal self — and, as we have seen, we are sometimes justified in making the latter kind of choice, tragic as it may be. But the point here is this: even if this person chooses to get a divorce, we can easily imagine her doing so without *blaming* her husband's normal self for what happened. Indeed, if anything, she might *pity* her husband's normal self for having to share a body and brain with a cheating self that will make it hard for anyone to want to date him in the future.

This way of thinking about responsibility is even more apparent when we turn to other, more obvious cases of psychological change. Consider, for example, a man who suffers from depression. His "normal self" does everything he can to keep his "depressed self" at bay: he takes medication, sees a therapist, studiously avoids anything that might trigger his depression, and makes sure that everyone in his life knows that he might, despite all these efforts, succumb to depression for days or weeks or months. Then, one day, he succumbs to depression and misses several important meetings at work. Now suppose that his normal self comes back out a few days later, tells his boss what happened, apologizes profusely, and promises to do everything he can to make it up to the company and never let it happen again. In this case, we can easily imagine this man's boss choosing to fire him — since, after all, people have to be able to interact *as people* for purposes of employment. Still, even if this man's boss chooses to fire him, we can easily

imagine her doing so without *blaming* his normal self for what happened. Indeed, if anything, she might *pity* his normal self for having to share a body and brain with a depressed self that will make it hard for anyone to want to employ him in the future.

Of course, as I have said, philosophers have traditionally thought that agent identity is only *necessary* for moral responsibility, not that it is *sufficient* for moral responsibility. Thus, advocates of course-grained theories of agent identity could try to accommodate the conclusions I have drawn here by arguing that (a) agent identity is, in fact, present in these cases but that (b) other necessary conditions for moral responsibility are absent. For example, in the depression case, one could say that (a) the man's normal self is one and the same moral agent as his depressed self but that (b) the man's depressed self is not fully *autonomous* (where autonomy is a necessary condition for moral responsibility), and therefore (c) the man's normal self is not fully morally responsible for what his depressed self does. My own view is that these explanations will not take us very far, especially in more typical cases like the infidelity cases (at least not unless we invent many more necessary conditions for moral responsibility than we currently have). But what matters for our purposes here is that, whatever we say about traditional theories of agent identity, a time-slice theory of agent identity is able to explain and justify our intuitions about all of the cases we have considered here in a simple, unified way: it says that we are not praise- or blameworthy for any of our past actions, but we are still complicit in, criticizable in light of, and/or liable for most, if not all, of them – and that our degree of complicity, criticizability, and liability for our past actions depends at least in part on our degree of physical, psychological, and narrative continuity with our past selves.

4. Conclusion

I have argued that, even if we are not one and the same moral agent across time, we can still be morally responsible for what we did in the past in a variety of ways. The upshot is that, if we accept a fine-grained theory of agent identity (coupled with standard assumptions about moral responsibility), then we will not, as many philosophers have assumed, have to radically revise our considered judgments about personal responsibility. Instead, we will be able to *affirm* our considered judgments about personal responsibility – while simultaneously revealing new layers in these judgments which we can then affirm as well.

Of course, we would need to consider many other theories of agent identity, as well as many other theories of moral responsibility, before we can reach a conclusion about which theory of agent identity has the most plausible implications for personal responsibility all things considered. Nevertheless, if my arguments here are correct, then we should shift the burden of proof away from fine-grained theories of agent identity in general and towards course-grained theories of agent identity in general, at least with respect to this issue. After all, a good theory of personal responsibility needs to do more than explain and justify our practice of holding each other responsible for past actions. It also needs to explain and justify our practice of holding each other responsible for *different past actions in different ways*. I have argued that a time-slice theory of agent identity accomplishes these aims remarkably well: it vindicates our judgment that we are morally responsible for most, if not all, of our past actions as well as our judgment that physical, psychological, and narrative continuity play a central (but, it turns out, non-foundational) role in determining what we are responsible for. The question that we have to ask, then, is whether course-grained theories of agent identity accomplish these aims too. That is, can

a theory of agent identity that makes us directly praise- or blameworthy for *all of our past actions* (setting aside radical breaks in physical, psychological, and/or narrative continuity) really explain and justify the subtlety and variety of our everyday practice of praising and blaming each other for what we did in the past? If so, then fine- and course-grained theories of agent identity both have plausible implications for personal responsibility, and we can decide between them on other grounds. If not, then fine-grained theories of agent identity have a clear advantage over course-grained theories with respect to personal responsibility, and we should accept a fine-grained theory unless we have very strong independent reason not to.[8]

References

Bratman, Michael. 1999. *Faces of intention.* New York: Cambridge University Press.

Feinberg, Joel. 1984. *Harm to others*. New York: Oxford University Press).

Degrazia, David. 2005. *Human Identity and Bioethics.* Cambridge: Cambridge University Press.

Gilbert, Margaret. 1990. Walking together: a paradigmatic social phenomenon, *Midwest Studies in Philosophy* 15:1–14.

Glannon, Walter. 1998. Moral responsibility and personal identity, *American Philosophical Quarterly* 35: 231-49.

Hart, H.L.A. and A.M. Honoré. 1985. *Causation in the law*, Rev. 2nd ed. New York: Oxford University Press.

Hawley, Katherine. 2010. Temporal parts, *The Stanford encyclopedia of philosophy*. ed. Edward Zalta. URL = <http://plato.stanford.edu/archives/win2010/entries/temporal-parts/>.

Christopher Kutz. 2000. *Complicity: Ethics and Law for a Collective Age*. Cambridge: Cambridge University Press.

Maddell, Geoffrey. 1981. *The Identity of the Self*. Edinburgh: Edinburgh University Press.

[8] Acknowledgements.

Nagel, Thomas. 1979. *Mortal Questions*. New York: Cambridge University Press.

Olson, Eric. 1997. *The human animal: personal identity without psychology*. Oxford: Oxford University Press.

Parfit, Derek. 1984. *Reasons and persons*. Oxford: Oxford University Press.

Plato. 1997. The republic, in *Complete works*, ed. John Cooper. Indianapolis: Hackett Publishing Company, pp. 971-1223.

Schechtman, Marya. 1996. *The constitution of selves*. Ithaca: Cornell University Press.

Shoemaker, David. 2012. Responsibility without identity, *The Harvard Review of Philosophy* XVIII: 109-32.

Shue, Henry. 1999. Global environment and international inequality, *International affairs* 75: 3, pp. 531-545.

Strawson, Galen. 2009. *Selves*. Oxford: Oxford University Press.

Velleman, J. David. 1997. How to share an intention, *Philosophy and phenomenological research* 57, pp. 29–50.

Williams, Bernard. 1981. *Moral Luck*. Cambridge: Cambridge University Press.