Altshuler, Roman (Kutztown University):

Ownership, Autonomous Agency, and Diachronic Choice

A great deal of recent philosophical writing explores the conditions for something like responsible or autonomous action.[1] The question is, essentially, a question of ownership: aside from an agent's simply wanting or choosing to φ, what else must hold in order for the agent's φ-ing to be *her own*, in a sense that is not vitiated by manipulation, upbringing, or other such factors? My aim here is to examine ownership for choices of a specific kind: diachronic choices. Unlike synchronic choices—like a choice to buy the blue or the brown socks, which is an event that can be pinpointed in time—diachronic choices are not made *at* a given time, but rather displayed in the synchronic choices made on their basis. The professor who consistently decides to meticulously grade her students' papers seems to be exhibiting, in each of these synchronic choices, an underlying choice. I will first argue that standard approaches to autonomous and responsible agency fail because they cannot account for ownership. Second, I argue that accounting for ownership requires a notion of diachronic choice, whereas the standard approaches draw either on atemporal features of agency or merely synchronic choices. Finally, I will defend an existential model of diachronic choice that can ground ownership.

The "something extra" needed for ownership has been spelled out in a number of ways. Real-Self views (Frankfurt) posit some second-order proattitude toward the agent's volition; a Whole-Self view (Arpaly and Schroeder) may instead argue that the volition must be integrated into the agent's psyche. Such views, however, do not fully explain ownership of the added condition: why should higher-order volitions, or even integration, make a difference to ownership unless those conditions are themselves owned? Since these conditions are not chosen, they appear to lack ownership. A further approach (Korsgaard and Bratman) takes autonomy to require acting in accordance with one's diachronic principles or plans, which confer ownership insofar as the agent chooses and identifies with them. But this seems problematic as well: the choices of principles or plans are revisable, and seem as synchronic as any choices made on their basis. If so, how can they automatically carry an ownership they can in turn bestow on other choices?

Fischer and Ravizza propose that we are responsible when we act on moderately reasons-responsive mechanisms that are themselves owned. Ownership in turn, requires that agents take responsibility for their mechanisms before they can be responsible for acting on them, and taking responsibility is a *process* involving three components. An agent has taken responsibility when he (1) sees himself as an agent, (2) sees himself as a fair target of reactive attitudes on the basis of the exercise of that agency, and (3) this self-conception is based on appropriate evidence. But there are problems with this view as well: It explains why we have no grounds to complain when we are held responsible, but

[1] There may be distinctions worth drawing between these concepts (Arpaly, for example, argues that autonomy is not required for responsibility, yet some features of her account may seem more pertinent to questions of autonomy than responsibility). Moreover, other authors use different terminology: Velleman speaks of full-blooded action, while Bratman emphasizes self-government. I will ignore these distinctions for the purposes of this paper.

not why we *are* responsible. And since taking responsibility on this view is a process in the *past*, it is not clear why we continue to be responsible for our mechanisms in the future, where unexpected situations may arise.

Such accounts, in other words, fail to explain ownership for autonomous or responsible action because they either (1) ground responsibility in some feature of agency that is not chosen, (2) allow for sudden reversal of that feature, thereby undermining ownership of it, or (3) assume that we can retain responsibility on the basis of past choices despite finding ourselves in novel situations that could not have been foreseen when our choices were initially made. Grounding responsible agency in diachronic choice avoids all three pitfalls by allowing us to hold that the ownership- conferring features of agency are chosen, persisting, and constantly renewed.

I will attempt to clearly articulate the existential alternative: when we act, our actions are expressive of self-defining projects. These projects, however, are not ones we choose synchronically. Rather, we discover that we have *always already* chosen them. That is: there is no moment of choice in which we adopt them, but rather we find the choices echoed in our acting on them. Moreover, since what we have already chosen is expressed in what we do, the content of the prior (and ongoing) diachronic choice depends not on some past volition, but on our continuously acting out the projects that define our practical identities. The ownership is not conferred by a synchronic act, and thus does not require a further account of how that act itself can be owned. And it is continuously clarified (and revised) in light of our ongoing self-identification, thus allowing for ongoing responsibility. My aim will be to explain and defend this model as a viable alternative to the existing views.

Amaya, Santiago (Universidad de los Andes):

Out of Habit

On January 1st 2009 police officer Johannes Mehserle shot Oscar Grant dead on a BART platform in Oakland, California. Mehserle didn't intend to shoot Grant. Yet, after warning him several times that he was going to be tased unless he stopped resisting arrest, the officer drew his gun and shot him once. Mehserle meant to reach for the taser, but instead reached for his gun.

During the trial the prosecution insisted the killing was intended. It was unreasonable to suppose that Mehserle had confused his gun for his taser. Both looked very differently and were carried on opposite sides of his belt: the gun on his right and the taser on his left. The defense counter-argued that the confusion was the only explanation consistent with Mehserle's behavior before and after the shooting. According to witnesses, immediately after it happened, he took his hands to his head and exclaimed with disbelief, "Oh shit, oh shit, I shot him."

Many philosophers think that beliefs play a crucial role in the genesis of action. If

someone acts with an intention, then the person's action is guided by her beliefs. The beliefs guide the action by directing the agent towards ways of achieving what she intends. In this respect, they help rationalize it. They reveal the connection the agent saw between her behavior and what she intended with it.

Slips, however, are a challenge to this view. Even though Mehserle did not intend to shoot Grant, he deployed his weapon with the intention of subduing him. Yet, his confusion cannot be explained attributing him false or otherwise muddled beliefs, say, about the location of his taser. After years of police experience, Mehserle had internalized the routine of reaching for his right holster when subduing a suspect who appeared threating to him. It was an old *habit*, not a *deviant belief* about how to execute his intention, what occasioned the slip.

It is tempting to try to resist this challenge in two different ways. One might argue that the slip is not properly an instance of a person acting with an intention. Although it can seemingly be described as such, in reality it is just a case of a person who fails to act on the intention she has in mind. Alternatively, one might argue that even though in the slip the action is not in line with what the person knows, it is in line with her beliefs—her beliefs at the time just so happened *no*t to be in line with what she knows.

In the paper, I discuss various ways of developing these responses and show why they are not as promising as they might initially seem. *Non-agentic responses*, I argue, tend to mischaracterize the mistake. For instance, discounting slips as instances of deviant causation overlooks how rational the mistake can be. In it, the behavior is not a brute response to the intention. Quite the opposite: it is semantically related to the agent's intention and would have been a good implementation of it in nearby possible worlds. In other words, rather than being a brute response, it is a *rational approximation* to the correct way of implementing the agent's intention.

On the other hand, *doxastic forms* of resisting the challenge overlook the ways in which habits can diverge from beliefs. Even though both are dispositions that guide behavior, I argue that they have a different functional profile. Slips do not spill across situations, which suggests that the habits involved in them are not *inferentially integrated* as beliefs tend to be. They are also eminently revisable mistakes, which suggests that habits need not be, as beliefs are, *attitudes of endorsement*. Thus, although the mistakes are attributable to the agent, they cannot be traced to a false or otherwise muddled belief in her.

The picture that emerges is a positive one. Despite being mistakes, slips are genuine instances of agency. Hence, they are candidates for attributions of responsibility. In them, the agent executes an intention. The execution is rational in the light of the agent's acquired habits. And the habits guide what the agent does. As such, slips provide a model for thinking about the role that habits can play in everyday responsible action. According to it, whereas habits are not reducible to complexes of (instrumental) beliefs, they can serve as *defaults for the execution of the agent's intentions*. In this capacity, they can play the guiding and rationalizing role that is customary attributed to beliefs.

Bazargan-Forward, Saba (University of California, San Diego):

Responsibility and Intervening Agency

When one person (P1) does something not in itself wrong but which enables another person (P2) to commit a wrong, that wrong is agentially mediated. Here is an example.

'Bulldozer' The employee of a construction company (P1) negligently leaves keys in a bulldozer overnight. The risk that makes this negligent manifests; a local vandal (P2) gains unauthorized access to the bulldozer, taking it for a joy-ride recklessly causing property damage.

In this example P1 plays an indispensable role in causing the property damage in that were it not for P1's negligence, P2 would not have been able to steal the bulldozer. But the damage was committed by P2 – not by P1. We can say that the harm which P1 brought about was "agentially mediated" by P2, and that P2 was an "intervening agent". In this case, P1 did not know that what he was doing risked unauthorized access to the bulldozer – be he was in a position to recognize that his conduct was risk-imposing.

Bulldozer is an example of agential mediation, which can be described more generally as follows. A harm φ is agentially mediated by P2 with respect to P1, act A1, act A2, for times t1 and t2, iff:

1

. (i) P1 commits A1 at t1, which significantly increases the probability that P2 will voluntarily commit A2 at t2, which causes a harm φ.

. (ii) P2 would not have committed A2, and φ would not have occurred, if P1 had not committed A1.

. (iii) P1 knows (or is in a position to know) both of the above.

To avoid confounding variables, it is important to keep morally relevant features constant between P1 and P2. I will assume that though P1 intends A1 and P2 intends A2, neither intends φ. I will also assume that P1 is in a position to recognize that committing A1 imposes a risk of φ's occurrence via P2's agency; likewise, P2 is in a position to recognize that committing A2 imposes a risk of φ's occurrence.

Given this account of intervening agency, is agential mediation relevant to how we should assess P1's and P2's conduct? If so, in what way? The existing literature on the moral relevance of intervening agency is dominated by a narrower version of that question: does P2's status as an intervening agent diminishes P1's responsibility for ϕ?[1] A prominent pre-theoretic view is that when the upshot of one's action 'passes through' the agency of another, one's responsibility for the upshot is diminished (or even

eliminated) – as if the other person were like a resistor in an electrical circuit.[2] The focus, though, on whether intervening agency should affect our assessment of P1's conduct has obscured a distinct and equally important issue which has gone

[1] See, for example (Hart & Honoré, 1958), (Kadish S. , 1985), (Zimmerman, 1985), (Hurd, 2001), and (Moore, 2009). [2] See for example what in legal theory is known as the "Voluntary Intervention Principle", versions of which are accepted in Anglo-American criminal and tort law. Hart and Honoré famously write that "[t]he free, deliberate and informed intervention of a second person, not acting in concert with the first, and intending to bring about the harm which in fact occurs or recklessly courting it, is normally held to relieve the first actor of criminal responsibility" (Hart & Honoré, 1958, pp. 42-3). They attempt to defend this view by appealing to an agent-causal view of human action. (See also (Kadish S. , 1985, p. 371) in which he argues that in criminal law "voluntary actions cannot be said to be caused"). I will not comment on the (im)plausibility of this defense, as it has been sufficiently criticized by others. See especially (Moore, 2000).

. unnoticed: whether intervening agency should affect our assessment of P2's conduct. This is the question I will be focusing on here. That is, I will not be investigating whether P2's status as an intervening agent makes P1's conduct less bad. Instead, I will turn this issue on its head by investigating whether P2's status as an intervening agent makes P2's conduct more bad. And I will argue that it does: holding all else fixed, P2 is responsible for two wrongs whereas P1 is responsible for at most one. More specifically, I will argue that though each of P1 and P2 bears some responsibility for ϕ, P2 wrongs P1 by having wrongfully made it so that ϕ is a consequence of what P1 does. Put differently, P1 has a defeasible claim against P2 that P2 refrain from acting in a way that makes what P1 does morally worse. I will then argue that this has important implications for the comparative liability of P1 and P2 for ϕ, which in turn has consequences for compensatory and defensive liability in cases where compensatory or defensive costs have to fall on either P1 or P2. The upshot is that what P2 does in Bulldozer is worse than what P1 does, not just because P2 imposes a greater risk of causing a harm ϕ, but because P2 wrongs P1. I will argue that this lesson can be generalized to all cases of intervening agency where neither P1 nor P2 intend the harm ϕ.

Works Cited

Hart, H. L., & Honoré, T. (1958). Causation in the Law (2nd ed.). Oxford: Clarendon Press.

Hurd, H. M. (2001). Is it Wrong to Do Right When Others Do Wrong? Legal Theory, 7(3), 307-340.

Kadish, S. (1985). Complicity, Cause and Blame. California Law Review, 73, 323-410.

Moore, M. S. (2000). The Metaphysics of Causal Intervention. California Law Review, 88, 827-878.

Moore, M. S. (2009). Casuation and Responsiblity. New York: Oxford Universirt Press.

Zimmerman, M. J. (1985). Intervening Agents and Moral Responsibility. The Philosophical Quarterly, 35(141), 347-358.

Beglin, David (University of California, Riverside) Abstract for "The Nature of Blame and Our Reasons for Forgiveness"

When ought we exempt certain agents from our practices of holding responsible? Call this the question of exemption. Many philosophers have grappled with this question, no doubt for a variety of good reasons. Recently, one group of philosophers has developed a particular strategy for answering it. To answer the question of exemption, these philosophers hold, we must first consider the nature of blame. I'm sympathetic with this approach. I'm less sympathetic, however, with a particular view that employs it. According to the Communication View, blame is a form of moral address, communicating a message to its object, and whether someone is a morally responsible agent depends, at least in part, on whether blame's distinctive communication can be intelligibly aimed at that person. This view has it, then, that we ought to exempt those agents who aren't intelligible targets of blame's moral address. I doubt the Communication View, because I doubt that blame is fundamentally about communication. Particularly, I suggest that the aspects of blame that are relevant to the question of exemption are those aspects that ultimately factor into why we blame people in the first place, and I doubt that blame's communicative aspect ultimately factors into why we blame people. To make my case, I turn to forgiveness. I argue that forgiveness is conceptually connected to blame: why we forgive people seems to depend, like exemption, on why we blame people in the first place. And it seems that when communicating messages factors into our reasons for forgiveness, it does so only due to our concern that people show us (or others) proper regard or goodwill. I suspect, then, that we ultimately blame people out of this concern, and so I suspect that it is this concern that is the key to answering the question of exemption, not blame's communicative aspect.

Bonicalzi, Sofia (University College London): Identification and rationality. Looking for a middle path between internalism and externalism about responsibility

During the last few decades, the cartography of moral responsibility theories has become more and more tangled. Judging from the huge literature on the topic, one of the most thought-provoking lines in the debate is represented by the growing family of compatibilist actual-sequence accounts. According to such views, given a deterministic

scenario, the access to alternative possibilities for action does not represent a necessary prerequisite for moral responsibility. In the absence of a reference to alternative possibilities, different characterisations of the underpinnings of moral responsibility have been outlined. However, the most relevant divide is likely to be the one between *internalist* and *externalist* actual-sequence views.

Internalism – usually associated with *self-disclosure* and *real-self views* – is committed to the claim that one is morally responsible for something if she is able to identify with the mental states determining the occurrence of the action, no matter their origin and nature (Frankfurt 1971; McKenna 2012). In turn, externalist accounts ground moral responsibility on elements that are located beyond the structure of the individual will, including one's ability to respond to intersubjectively recognised normative constraints (Wolf 1990; Nelkin 2011). In *historical* externalist accounts, one being morally responsible for something also depends on one's action displaying a causal history devoid of episodes of hidden or uncovered manipulation (Fischer and Ravizza 1998).

Both internalist and externalist accounts appear to have considerable drawbacks. Internalism has a hard time discriminating between responsible (MRs) and non-responsible (NMRs) scenarios in borderline situations, including cases of weakness of the will, out of character acts, manipulation, or deviation from the standards of rationality (see Mele 1995; Wolf 1987). Externalism, in particular in its historical version, has been accused of introducing arbitrary distinctions between different causal histories and aetiologies of actions, which could be ultimately incompatible with the claim that determinism does not undermine responsibility (see Double 1991; Pereboom 2014).

Aiming to address some of those problematic issues, I propose a compatibilist-friendly account, which combines internalist and externalist insights, and defend the claim that a kind of self- disclosure view, adequately supported by an account of normative competence, could give reason of many of our usages of the concept of moral responsibility.

Following internalism, the distinction between MRs and NMRs is primarily grounded on the characteristics of the internal structure of the decisional process leading to action. Drawing on Watsons's *attributability* (Watson 1996), one must distinguish between cases in which the action reveals something morally relevant about the one who makes it, and situations where it does not (the latter possibly including cases of manipulation, compulsion, and resultant luck).

The leading intuition would be that, in MRs, one's motivational structure is *explanatory relevant* with respect to the action (Björnsson and Persson 2012), in a way that makes the agent able to identify with it. However, differently from classic internalist real-self views, according to the view I favour identification does not concern individual mental states: recognising the action as a final step in the deliberative process she goes through, one identifies with the result of the global choice- making process.

Identification-based attributability does not exhaust the required conditions for moral responsibility. Externalist integrations are needed in order to provide a functioning

framework, able to deal with cases of irrationality or extreme distance from moral standards. Paraphrasing Haji (2012), such requirement is to be understood as responsibility's *debt* to rationality: moral responsibility requires rational control, which is meant to be an externalist component, independent of one's identification with the process of choice leading to action. For the condition of normative

control to be satisfied, the agent should be able to accept intersubjectively recognised reasons for action as compelling motives.

In this sense, a gradualist, rather than a *black & white*, conception of moral responsibility is to be preferred: one's degree of moral responsibility might vary according to her ability to consider the reasons for action that are in place in a given context. Differently from traditional historical externalism, it is not one's unfortunate causal history that would reduce (or eliminate) moral responsibility, which would rather depend on the way in which one's reasoning abilities are arranged, given the experiences she underwent. Thus, the relevant demarcation line is not linked either with the strength of one's prevailing motives or with the quality of one's preferences, deriving instead from the possibility for the agent to correctly exercise some sort of moral reasoning.

To sum up, claiming that X is morally responsible for the action Y would imply that (1) the action Y reveals something morally relevant about X (internalist constraint) and (2) X is able to grasp the moral meaning involved in the action Y (externalist constraint). As a result, moral responsibility attributions would be understood as explanatory tools, whose role is to portray the degree at which one's choices and actions are to be referred, and properly attributable, to a reasonable decision-maker.

References

Björnsson G., K. Persson (2012), "The explanatory component of moral responsibility", *Noûs*, vol. 46, no. 2, pp. 326- 354;
Double, R. (1991), *The non-reality of free will*, Oxford: Oxford University Press;
Fischer, J. M., M. Ravizza (1998), *Responsibility and control: a theory of moral responsibility*, New York: Cambridge University Press;
Frankfurt, H. G. (1971), "Freedom of the will and the concept of a person", in Id., 1988, pp. 11-25;
Haji, I. (2012), *Reason's debt to freedom: normative appraisals, reasons, and free will*, New York: Oxford University Press;
McKenna, M. (2012), "Moral responsibility, manipulation arguments, and history: assessing the resilience of nonhistorical compatibilism", *Journal of Ethics*, vol. 16, no. 2, pp. 145-174;
Mele, A.R. (1995), *Autonomous agents*, New York: Oxford University Press;
Nelkin, D. K. (2011), *Making sense of freedom and responsibility*, New York: Oxford University Press;
Pereboom, D. (2014), *Free will, agency and meaning in life*, Oxford: Oxford University Press;
Watson, G. (1996), "Two faces of responsibility", *Philosophical Topics*, vol. 24, no. 2, pp. 227-248;

Wolf, S. (1990), *Freedom within reason*, New York: Oxford University Press;
Wolf, S. (1987), "Sanity and the metaphysics of responsibility", in F. Schoeman, ed., 1987, *Responsibility, character, and the emotions*, New York: Cambridge University Press, pp. 46-62.

Brandenburg, Daphne (Radboud University): The Nurturing Stance

Psychiatric practice provides a fascinating challenge to a family of theories of moral responsibility. Clinicians report they often hold people responsible for norm-transgressions even though blame is considered *inappropriate*. This presents a conundrum to those theories that, following P.F. Strawson, define an agent's responsibility in terms of being an appropriate target of praising and blaming attitudes.

On the basis of these insights from psychiatry, Hannah Pickard has recently argued against the link between the capacity to meet shared norms and demands on the one hand, and being the appropriate target of praise and blame on the other hand. According to her responsibility should be attached to the normative capacities that a person has, but *detached* from moral praise and (especially) blame. (Pickard, 2013)

In reply to Pickard I argue that there are two different solutions available to the conundrum. Both solutions qualify the link between responsibility, and blame in interesting ways, but ultimately do not completely detach one from the other. The first solution has been argued for by others and connects to a currently thriving debate. (Smith, 2007; Coates & Tognazzini, 2013; Todd, 2012; King, 2015) It qualifies the link between being responsible and blameworthy on the one hand, and the appropriateness of expressing blaming attitudes on the other hand.

I focus on a second solution which, I argue, provides a novel contribution to the literature. This solution qualifies the link between the incapacity to meet a norm or demand on the one hand, and exemptions from responsibility on the other. I will argue that psychiatric practices provides us with a form of holding responsible that does not track full-blown capacities but tracks potential capacities instead. And because it has another target it renders a blame response inappropriate. This form of holding responsible is an important addition to the Strawsonian theories at stake and generalizes to practices outside of the clinic too.

Coates, D. J., & Tognazzini, N. A. (2013). The Contours of Blame. In D. J. Coates & N. A. Tognazzini (Eds.), *Blame: Its Nature and Norms* (pp. 3–26). Oxford University Press.

King, M. (2015). Manipulation Arguments and the Standing to Blame. *Journal of Ethics and Social Philosophy*, *9*(1), 1–20.

Pickard, H. (2013). Responsibility without blame: Philosophical reflections on clinical practice. *Oxford Handbook of Philosophy of Psychiatry*, 1134–1154.

Smith, A. M. (2007). On Being Responsible and Holding Responsible. *Journal of Ethics*, *11*(4), 465–484.

Todd, P. (2012). Manipulation and Moral Standing: An Argument for Incompatibilism. *Philosophers' Imprint*, *12*(7).

Bruno, Daniele (Humboldt University, Berlin): Must We Worry About Epistemic Shirkers?

Abstract for the Gothenburg Responsibility Project Conference #1, 24-27 August 2016

It is generally acknowledged that ignorance of the import of one's actions can at least sometimes excuse one from actions that would under normal circumstances be morally reprehensible. This idea is reflected in the following plausible principle:

> *Knowability Condition:* For all obligations x, S is blameworthy for violating x only if S can know she has x.

However, subscribing to the Knowability Condition thus phrased might also give rise to a worry: if blameworthiness is irretrievably linked with knowability, what about agents who consciously attempt to make epistemic access to some morally relevant fact impossible for themselves? At first sight, doing so might seem to afford them the opportunity to actively create excuses for their actions, thereby evading blameworthiness for their moral failings.

In this paper, I have a closer look at this phenomenon, which we may call *epistemic shirking*. As I shall understand it, an agent S shirks on an obligation to φ iff S attempts to evade blameworthiness for φing at t_2 by performing an action A at t_1, which causes S to not meet the necessary conditions for being blameworthy for φ-ing at t_2. Epistemic shirkers do so by specifically impairing their ability to know morally relevant propositions. Examples of such behaviour abound. Think of an agent who turns off a TV programme about the conditions animals suffer in factory farms, for no other reason as to not be exposed to the potentially morally significant information contained therein.

It has been argued (by Sorensen 1995 and Wieland 2015) that the possibility of pre-emptive shirking leads to an unfortunate regress problem when looking for potential grounds on which to hold epistemic shirkers blameworthy for their behaviour. In short, the problem is that if we point to a specific moral demand that the shirker fails to meet (such as "she ought not to have made her moral obligation unknowable to herself"), this opens up the possibility for the shirker to also make this further moral demand unknowable, and so on for any further higher-order obligations we might make reference to.

After laying out the regress problem in detail, I discuss a solution to the problem that was recently put forward by Wieland (2015). Wieland suggests that shirking is blameworthy because shirkers are subject to a self-referential obligation that is itself not shirkable.

> *General Law:* I should refrain from making or keeping my obligations unknowable [including this very obligation]. (cf. Wieland, 2015, pp. 297f.)

I argue that this putative obligation will not solve the problem in a satisfactory manner, as Wieland faces a dilemma when spelling out exactly what it means to keep one's obligations unknowable in the context of General Law. On the one hand, a strong interpretation leads to implausibly demanding moral requirements, while on the other hand, a weaker interpretation fails to secure blameworthiness for all shirkers unless it can be shown to be necessarily knowable for all moral agents.

Does this mean that shirking is actually a winning strategy? In the last part of the paper, I argue that this is not so. I lay out in detail how a careful re-evaluation of shirkers' reasons and motivations for their decision to shirk can lead us to sufficient grounds for a solid ascription of blame in all the cases we should want it to do so. I show that the very idea of shirking on an obligation implies a blameworthy disrespect of the regarding obligation on the part of the shirker – specifically given her *risk* of breaching it: for someone to be rightfully called a shirker, the agent has to be aware of her having a (possibly as of yet not fully defined) obligation x, or at most suspending judgement on this matter. Shirking is a *reaction* to this possibility occurring to her. Thus, in every instance of epistemic shirking, there is a risk-based moral obligation that the shirker consciously and impermissibly disregards. I argue that the principle that we ought to avoid risking violation of our obligations follows easily from a natural understanding of morality and is thus accessible to any normal agent capable of entertaining moral considerations (as shirkers are, per definitionem), implying it cannot be pre-emptively shirked. Admittedly, the risk-based solution does not return as blameworthy agents with certain very skewed moral upbringings, or agents who chose to deteriorate their epistemic standing for non-moral reasons absent any evidence of moral risk for these actions. This however, is, I take it, a result that correctly reflects our intuitive blameworthiness-judgements about these characters. In the cases that lead to the basic worry that drives the discussion – shirking by ordinary agents as defined above – the risk-based considerations provide solid grounds for the ascription of blame.

Clancy, Sean (Syracuse University):

Psychopaths, Ill-Will, and the Wrong-Making Features of Actions

Are psychopaths morally blameworthy for their bad actions? Many recent treatments of this question have focused on psychopaths' capacity, or lack thereof, to express *ill-will* by acting. Psychopaths apparently understand that others can be harmed, and thus their actions can express the judgment that the harm they inflict on others is unimportant. But psychopaths apparently do not understand that others have moral standing, and thus their actions cannot express contempt for other persons *qua* moral patients. While the parties to the ill-will debate agree as to which attitudes psychopaths can express, they disagree as which attitudes count as ill-will. The debate seems to have reached an impasse.

I argue here that this impasse reflects an implicit disagreement as to which features of actions are wrong-making. Ill-will is best understood – and seems to be understood by the participants in the debate over psychopathy – as an objectionable attitude towards the features of actions which make them wrong. As I hope to show, the

question of precisely which features are wrong-making is more difficult than has previously been appreciated. Even when a given set of normative assumptions is shared, it is possible for different parties to reasonably identify different features of an action as wrong-making.

This, I contend, is the root cause of the disagreement over the ill-will question. The expression of ill-will requires awareness of the wrong-making features, and psychopaths are aware of some features but not others. Some parties implicitly identify the wrong-making features as those of which the psychopath *is* aware – such as the harmfulness of his actions – and conclude that he *does* express ill-will. Others implicitly identify the wrong-making features as those of which the psychopath is *not* aware – such as the fact that he harms *persons* with moral standing – and conclude that he does *not* express ill-will. This underlying disagreement explains why the ill-will question has reached its current impasse; exposing the disagreement reveals a way in which the impasse can be broken.

Clarke, Randolph (Florida State University): Moral Responsibility, Guilt, and Retributivism

Abstract: This paper defends a minimal desert thesis, according to which someone who is blameworthy for something deserves to feel guilty, to the right extent, at the right time, because of her culpability. The sentiment or emotion of guilt includes a thought that one is blameworthy for something as well as an unpleasant affect. Feeling guilty is not a matter of inflicting suffering on oneself, and it need not involve any thought that one deserves to suffer. The desert of a feeling of guilt is a kind of moral propriety of that response, and it is a matter of justice. If the minimal desert thesis is correct, then it is in some respect good that one who is blameworthy feel guilty—there is some justice in that state of affairs. But if retributivism concerns the justification of punishment, the minimal desert thesis is not retributivist. Its plausibility nevertheless raises doubt about whether, as some have argued, there are senses of moral responsibility that are not desert-entailing.

Longer Summary:

Let us say that a guilty person is someone who is morally blameworthy for something, where the blameworthiness at issue is a mode of moral responsibility. With a guilty person so understood, does a guilty person deserve anything bad because of, or in virtue of, her guilt?

This paper defends the following minimal desert thesis: a guilty person deserves to feel guilty, to the right extent, at the right time, regarding the thing for which she is to blame, and deserves this because of her culpability. To feel guilty is unpleasant; it is bad to some degree. The guilty person thus deserves something bad.

The paper examines a recent objection from Derk Pereboom to the minimal desert thesis. The objection, I argue, stems from a misconstrual of the thesis. I then identify an apparent implication of the thesis, one that T. M. Scanlon, despite accepting the thesis, appears reluctant to accept. I point out that the thesis is not a form of retributivism, as this is standardly understood. Nevertheless, if the minimal desert thesis is correct, then moral responsibility bears an intimate connection to desert; it is, as some have said, basic-desert-entailing.

The first section of the paper sets out a view of the sentiment or emotion of guilt. I take it for granted that some thought is partly constitutive of this state, and I defend the view that the thought is that one is blameworthy for something. Other candidates for the constitutive thought are rejected as more or less than what is required if a feeling of guilt is to be warranted. Of course, the feeling of guilt isn't just a thought; it includes as well an unpleasant affect. It feels bad to feel guilty.

Section two examines the connection between the constitutive thought and the unpleasant affect of the feeling of guilt. The thought is part of what explains the affect. Here I defend the minimal desert thesis from an objection that Derk Pereboom has raised against it.

Pereboom cites a passage from Hilary Bok pointing out that the connection in question is disanalogous to that between a criminal conviction and the imposition of punishment. It is not, Bok says, that one imposes suffering on oneself because one thinks that one deserves to be made to suffer. Rather, the connection is like that between the recognition that one has lost a loving relationship and a feeling of heartbreak. The pained feeling is simply an appreciation of what one recognizes.

The minimal desert thesis, I point out, is entirely consistent with Bok's observation. The thesis does not concern the explanation of the pain of feeling guilty. It does not say that the unpleasant affect stems from a belief that one deserves to suffer. Rather, it offers a characterization of the normative status of this attitude: such a feeling by one who is blameworthy, the thesis says, is deserved. Hence, the point about explanation is no objection to the minimal desert thesis.

Section three offers a view, drawn from Joel Feinberg, of personal desert as kind of moral propriety, and a consideration of justice. The desert of a certain response is a *pro tanto* consideration favoring that response, one that can be outweighed by others. Thus, a response can be all-things-considered wrong even if deserved.

Feinberg suggests that personal desert is, or is like, a kind of fittingness between someone's actions or qualities and some responsive attitude. Such attitudes, he suggests, are the basic things that persons deserve; modes of treatment are deserved insofar as they express morally fitting attitudes.

It is important to say here that the fittingness of some attitude comes to desert only when it is a consideration of justice. The proviso is needed to explain why the unpleasant agent-regret of Bernard Williams's unlucky truck driver, although it might be fitting, is not deserved.

Still, a claim about the fittingness of this or that attitude is in better shape if it has something more than unstructured intuition to back it up. I draw from some ideas advanced by T. M. Scanlon to support the minimal desert thesis. Respect for those who have been wronged provides a reason for the guilty party to feel guilty. And she does not owe it to herself not to feel bad in this way when she is blameworthy.

In section four, I suggest that if a guilty person deserves to feel guilty, then it is in some way good that she do so. That state of affairs is in one respect just, in that the agent has a feeling that she deserves to have. And a state of affairs that is in some respect just is in

some respect good. Then, since to feel guilty is to suffer in some way, it would be in some respect good that a guilty person suffer in this way.

In earlier work, Scanlon associated desert with the idea that it is good that people who have done wrong should suffer. He later came to affirm the desert by the guilty of certain responses, including the feeling of guilt, but he still rejects this older idea as morally repugnant. I point out that the new idea that he accepts appears to entail a restricted version of the old idea: that it is good, in some respect, that a wrongdoer suffer the feeling of guilt (to the right extent and at the right time).

'Retributivism' is commonly used to refer to views alleging the desert of modes of treatment, and particularly punishment. In section five, I point out that, on this understanding, the minimal desert thesis is not retributivist. For the thesis does not say that any mode of treatment is deserved.

A feeling of guilt is distinct from any action taken by any person—the guilty party or anyone else—to punish or inflict suffering on the blameworthy individual. Hence the desert of a feeling of guilt isn't the desert of being treated in any such way.

I consider as well views from P. F. Strawson and Gideon Rosen on which reactive attitudes such as resentment, indignation, and the feeling of guilt are intimately tied to retributivism. As Rosen sees it, resentment and indignation are forms of anger. Each includes a desire that the perceived wrongdoer suffer in recognition of her wrongdoing, as well as the thought that this person deserves to suffer as a result of punitive sanction. I point out that this view does not apply to the feeling of guilt, since, as Rosen recognizes, the latter does not involve anger and need not be associated with a desire to suffer. A better candidate for the constitutive thought of the feeling of guilt, again, is that one is blameworthy for something. This thought is not explicitly retributive.

The final section takes up the notion of basic desert that Derk Pereboom has explicated. The basic desert of some response to some person is a moral propriety of that response, a propriety that obtains simply in virtue of some act or quality of that person. It is non-consequentialist and not a matter of contractualist considerations favoring any practices or institutions. The desert with which the minimal desert thesis is concerned is evidently basic desert. Thus, if the thesis is correct, then moral responsibility is basic-desert-entailing: there is at least one response, a feeling of guilt, that a blameworthy agent deserves simply in virtue of some act or quality of hers.

Pereboom thinks it unlikely that we are morally responsible in any basic-desert sense, but he proposes an alternative conception of moral responsibility that, he holds, applies to us. On this conception, too, it may be apt for a blameworthy agent to recognize her wrongdoing and feel bad about it. If so, I argue, then whatever forward-looking considerations might favor her having this attitude, some backward-looking consideration favors it, some consideration concerning the fact that she performed the act in question with such-and-such a quality of will. How is the propriety supplied by this backward-looking consideration supposed to be different from basic desert? Until Pereboom answers this question, I argue, he has not provided a viable conception of moral responsibility that avoids commitment to basic desert.

Moral Responsibility, Situationism, and Deery

Deery, Oisín (University of Arizona):

Implicit Bias

In this paper, I apply an appealing position about natural kinds in science and everyday life— the homeostatic-property-cluster (HPC-kind) view (e.g., Boyd 1988)—to debates about moral responsibility. According to this view, the choices for which people are morally responsible count as a natural kind just in case: (A) paradigmatic instances of such choices share a homeostatic cluster of control capacities in common, and (B) this cluster provides the best available support for our inductive and causal-explanatory generalizations in appealing to the kind. The view that emerges avoids a number of threats to moral responsibility. In this paper, I focus on the threats posed by situationism and implicit bias.

There is evidence that at least some choices for which we hold agents responsible might not be best explained by generalizations identified by a homeostatic cluster of capacities exercised by the agent, but instead by generalizations that appeal to external situational features. For instance, whether an agent chooses to help a bystander is sometimes best explained by whether there is noise nearby (Mathews and Canon 1975: 574–5). If there is, agents tend not to help; otherwise, they help. These findings threaten moral responsibility since they point to ways in which agents seem to lack control.

Yet even as such findings threaten moral responsibility, they also reveal how the sort of control that grounds responsibility might be enhanced. Once an agent becomes aware that her choices are liable to be influenced in certain ways by situational features, she can control for these influences—for example, by avoiding the relevant situations. In this way, a recovering gambling addict might foresee that if she goes to a friend's party, it will be difficult for her to avoid playing poker. So, she avoids the party altogether, thereby avoiding the foreseen difficulty. Indeed, findings in psychology indicate that the willpower the addict must exercise to avoid playing poker if she attends the party is a limited resource. Its depletion (by tiredness, overwork, and so on) will temporarily impair her ability to regulate her behavior in response to problematic impulses, once they have arisen (e.g., Baumeister and Tierney 2011).

Even non-addicted agents can employ such strategies, which Tamar Gendler (2015) calls strategies of "preemptive self-control," and which require effort to be exerted in anticipation of a problematic impulse, so as to avoid relying on willpower alone in resisting it (e.g., de Ridder et al. 2012). Preemptive self-control can be deployed to compensate for situational influences on choice, at least when an agent knows about them.

A useful framework for thinking about strategies of self-control is the "process model" developed by Angela Duckworth and colleagues (2015). On this model, an agent utilizes metacognition and prospection to prioritize strategies that have greater counterfactual reliability, in being more effective across a wider variety of situations, over methods with less counterfactual reliability, and to prioritize strategies that require fewer cognitive resources over ones that require more resources (2015: 206–11). The first strategy is

"situation selection," by which an agent consciously chooses to be in situations that (or with people whom) she foresees will maximize her self-control. Thus, a student might study in the library rather than at home, to avoid distractions. When a situation cannot be avoided, the next strategy is "situation modification." Thus, an agent might use a program that limits her internet access, to avoid distractions in meeting a work deadline. If that strategy fails in a situation, an agent can employ "attentional deployment," perhaps by counting backward from 100 during a heated dispute, to avoid escalating the conflict. When that fails, an agent can deploy "cognitive change"—for instance, by framing her boss's criticism of her as information rather than a measure of self-worth. Finally, when all else fails, agents can deploy willpower, or "response modulation," the most cognitively expensive and least reliable strategy of all. Yet even then, agents can notice which situations most often require the exercise of willpower, and they can try to avoid these situations in future.

I maintain that this framework also applies to known situational influences on choice. For instance, jurors who wish to deliberate fairly before delivering a verdict should avoid deliberating in a room where a trash can is overflowing with smelly pizza boxes, once they know that people make harsher moral judgments when exposed to such disgust-inducing stimuli (Wheatley and Haidt 2005). If the situation is unavoidable, the jurors might modify it, by requesting that the boxes be removed, or by deliberating as far as possible from them. Or they might redeploy attention, by looking away from the boxes or by eating mints to distract themselves from the smell, and so on.

The apparent problem for moral responsibility posed by situationist findings is thus defused by the HPC-kind view, due to the capacities for preemptive self-control that plausibly belong in the cluster of capacities that defines free and responsible choices, yet which do not feature in other accounts of free choice. Moreover, the HPC-kind view explains how agents can be responsible for choices they make that are influenced by implicit bias, by explaining how agents can control for biases in ways analogous to how they can control for situational influences (cf. Levy 2012).

References

Boyd. (1988). "How to Be a Moral Realist," in *Essays on Moral Realism*, Cornell UP: 181– 228.

Baumeister and Tierney. (2011). *Willpower: Rediscovering the Greatest Human Strength*, Penguin.

de Ridder et al. (2012). "Taking Stock of Self-Control," *Personality and Social Psychology Review*, 16:76–99.

Duckworth et al. (2014). "Self-Control in School-Age Children," *Educational Psychologist*, 49:199–217.

Gendler. (2015). Presidential Address to the Society for Philosophy and Psychology, Duke University.

Levy. (2012). “Consciousness, Implicit Attitudes, and Moral Responsibility,” *Noûs*, 48:21– 40.

Mathews and Cannon. (1975). “Environmental Noise Level as a Determinant of Helping Behavior,” *Journal of Personality and Social Psychology*, 32: 571–77.

Wheatley and Haidt. (2005). “Hypnotically Induced Disgust Makes Moral Judgments More Severe,” *Psychological Science*, 16: 780–84.

Gillespie, Laura (University of California, Los Angeles):

Between Friends: On the Possibility and Permissibility of Interpersonal Punishment in Relationships of Equality

There are a variety of familiar responses to wrongdoing in the human repertoire, the most controversial of which generally involve the imposition of some cost upon the wrongdoer. Two philosophical literatures concerned with this particular controversy are those on punishment and blame. In the first literature we encounter the many philosophical problems about punishment— problems both about what punishment is and about how it might be justified. The punishment literature has focused with near total exclusivity on the very real and pressing issues raised by institutional or state responses to wrongdoing. In the literature on blame, by contrast, the focus remains firmly in the domain of the interpersonal, but the sorts of responses to wrongdoing we attempt to understand and justify consist primarily in the holding of particular beliefs and attitudes, or the experience of so-called “moral emotions” such as anger. What has rarely, in either literature, been directly or sustainedly addresses is the question of what *action* we might be entitled to take toward or even against those who wrong us *interpersonally*. This is rich territory, including, potentially, much of the behavior common to interpersonal conflict, from the silent treatment to a cancelled plan to the slamming of a door. My project is to make a start at mapping this terrain, thinking both about the sorts of intentional actions that might be justified in response to interpersonal wrongdoing, and the sorts of reasons or motives for such actions that might stand to justify them.

In the paper I propose to present, I directly consider the nature and justification of interpersonal punishment, one particularly controversial form of intentional response to wrongdoing in the domain of the personal. In it I provide an answer to the question:

When and why might it ever be permissible to respond to the wrongdoing of our friends with the intentional imposition of a burden or deprivation? In this paper I aim to defend the position that (1) punishment is a relatively common, everyday form of treatment in the context of our friendships, and that (2) in certain kinds of cases such treatment will be not only permissible, but a requirement of the relationship. I aim, in other words, to defend the claim that you (probably) do and (probably) should sometimes punish your friends.

I begin by adopting a relatively capacious but principled understanding of punishment as that form of treatment which constitutes both a response to wrongdoing and the intentional deprivation or burdening of the wrongdoer. For it to be constitutive of some act that it so burdens or deprives the wrongdoer means that the burden or deprivation in question cannot be mere incident or accident. Consider such commonplace behavior as the silent treatment or the cancelling of plans in the wake of wrongdoing. These, along with the many other forms of withdrawal that tend to occur in the wake of wrongdoing and interpersonal conflict, constitute deprivations. Not all such deprivations are, of course, intentional in the relevant sense. We may withdraw in order to regain composure, to nurse our wounds, or simply to escape. These responses to wrongdoing, whatever deprivation they may constitute for the wrongdoer, are not instances of punishment. The deprivation in these cases is merely incidental. On other occasions, though, it will be constitutive of the wronged party's response that it constitutes a deprivation for the wrongdoer. I tell you, e.g., to go sleep on the couch not only to give myself some space, but for the reason that I will thereby deprive you temporarily of something you presumably value—your comfort, perhaps, or just the reassuring and intimate practice of sharing a bed. Here we have a form of interpersonal punishment, which, as such, stands in need of a justification.

To offer a satisfying defense of any such instance of punishment will, I argue, require that we appeal to a special class of reasons not easily categorized in terms of the standard distinctions between "backward" and "forward" looking reasons to blame or punish, and not for the kinds of reasons typically offered as potential justifications of punishment or of blame and blaming behaviors more generally. I do not defend interpersonal punishment on retributive grounds, nor do I defend it as a form of self- gratification, as a means of deterrence or behavior modification, as a requirement of self- respect on the part of the wronged party, or as a means of spurring appropriate guilt or remorse in the wrongdoer. I argue, rather, that it may sometimes be permissible to punish a friend *for the sake of the friendship*, where friendship is understood as a source of non- instrumental value for both parties. Where punishment is permissible, it will amount to a particular sort of deprivation that operates by way of a communicative capacity unique to that relationship. In such cases the deprivation in question will be of some common experience or expression of the value of the friendship itself, and this deprivation will work to communicate not merely by serving as a reliable or predictable means of bringing it about that the wrongdoer should come to have some set beliefs or attitudes that she formerly lacked. To be a friend, I will argue, is to bear a communicative capacity to make oneself understood by means of this sort of deprivation—a capacity that relies on a complimentary capacity to understand such deprivations as communicating (and as intended to communicate) information crucial to the friendship in which each has a share.

The aim of the proposed project is to advance discussions of both blame and punishment, by considering a set of potentially illuminating cases largely neglected in both literatures, and appealing to a potential justification that troubles the standard categorizations we use in thinking about what might justify either. In doing so, I aim to sketch a strategy of justification that avoids the most troubling aspects of "forward" and "backward" looking accounts alike, and provides a potential way in for those skeptical of either sort of account.

Gorman, August (Amanda) (University of Southern California): You don't seem like your Deep Self lately

An estimated 5% of people worldwide suffer from some form of clinical depression. Strikingly, it has been radically under-theorized by philosophers working on autonomous agency. The effects of the condition on the functioning of agency are well known to those who have suffered, making clinical depression arguably the most common cause of volitional disability. Yet little effort has been put towards understanding how we can accommodate the potential for the kind of agential threat that clinical depression poses on our best theories of autonomous agency. Much of the literature on impaired agency centers on cases in which a person lacks the ability to protect herself from errant urges or cases in which a person lacks the capacity to initiate crucial self-reflective agential processes, which has naturally lead to frameworks for thinking about autonomous agency that are designed with only the possibility of these sorts of impairments in mind. However, I argue that the proper understanding of clinical depression's threat to the will reveals a third way that agency can be undermined. In this talk I will sketch an explanation of the phenomenon and show how the explanation can draw on the basic framework of the hierarchical identification theory of autonomous agency, which was designed with these quite different cases in mind, given some simple modification to the framework. On my account, which I call the Volitional Gap Model, clinical depression undermines agency by divorcing a second-order volition to act on a desire to φ from an occurent desire to φ, but the possibility of this requires a new way of understanding volition. I will end by showing some of the advantages of adopting the Volitional Gap Model to explain the phenomenon of depression's impairment.

Gunnemyr, Mattias (Lund University) Participation as the Basis of Responsibility

In a globalized society we find ourselves connected to collective harms and wrongs, a connection that falls outside the concept of individual intentional wrongdoing. We might for instance buy a shirt made by workers working under slave-like conditions, work for a company that also produces landmines sold to some poverty-stricken country engaged in a military conflict, or participate in social-structural processes that makes single mothers vulnerable to homelessness. Traditional models of responsibility – such as the control principle or the individual difference principle – fails to explain the nature of our responsibility for these kinds of harms. I propose that the basis for holding someone responsible for a collective harm is whether this person has participated the social processes that bring this harm about. No participation without implication.

The currently most influential accounts of collective responsibility assume that there must be some collective act for which we assign responsibility for collective harms. However, this assumption prevents the possibility to assign responsibility for unstructured collective harm i.e. for collective harm that is not the result of a collective action such as global warming, overfishing or unjust social practices. I will argue that the accounts of collective responsibility proposed by Margaret Gilbert (2000, 2006, 2014),

Christopher Kutz (2000), Tracy Isaacs (2011) and Philip Pettit (2001, 2007) cannot explain responsibility for such harm. Agents who contribute to harm such as global warming aren't jointly committed and they lack participatory intentions. Additionally, these harms do not originate in the act of an organisation or a goal-directed collective. Nor do they originate in the collective decisions of an institutionalized decision-procedure. In addition, I will argue that Brian Lawson's (2013) modified version of Kutz's account of complicity partly shares this problem; it cannot explain responsibility for unstructured collective harm in cases where the agent unknowingly contributes to harm. I will also make the case that Michael Bratman's (1997, 2013) account of shared agency can not be used to explain responsibility for those harms, at least not without substantial revisions. For instance, agents who contribute to global warming do not share the intention to bring this harm about.

The idea of taking participation as the basis for responsibility originates at least partially in Iris Marion Young's (2011) social connection model for responsibility, and Maeve McKeown's (2015) interpretation thereof. Yet, in contrast to the accounts of collective responsibility and/or collective action already mentioned, the social connection model concerns primarily unstructured collective harms. I suggest that the common denominator for assigning backward-looking responsibility both for unstructured and for structured collective harm is whether you participate by your actions in the social processes that cause harm.

This account has at least two important advantages. First, it allows us to assign responsibility for unstructured collective harm as well as for structured collective harm since it requires participation in a social process instead of participation in a collective act. Second, taking participation as the basis of responsibility explains responsibility for harms I contribute to as well as responsibility for harms my actions depend on. Take the global garment industry as an example: When you buy clothes manufactured by workers working under slave-like conditions, you are responsible for the suffering of those workers. Yet, the responsibility cannot be understood correctly only considering what my act of buying the clothes will *contribute to*. The causal relation is reversed. The manufacturing of the clothes is a part of the social processes that make it possible for you to buy them. Your action *depends* on these harmful processes rather than contributes to them. Taking participation as the basis of responsibility still accounts for the familiar idea that you are responsible for harms that you contribute to. For instance, the money you pay for your clothes might contribute to the reproduction of social processes that most probably will result in workers being similarly exploited in the future. If that is the case, you have participated in social processes that cause harm.

There seems to be two obvious disadvantages with using participation as basis for responsibility. For one thing, an agent's participation in producing harm does not tell us anything about the agent's nature or degree of responsibility. This problem is however not particular to this approach. All accounts of the basis of responsibility must be complemented with an account of how to assign nature and degree of responsibility to those who are responsible. How the agent participates (intentionally, unintentionally; knowingly, unknowingly etcetera) and the nature of harm will most probably significantly influence the kind of response that is warranted. In some cases blame is appropriate, in some it is not. In some cases we are liable for the harm we have

participated in bringing about, in some we are instead responsible for ameliorating the social processes that causes these harms.
There is another worry. Since participation on my interpretation is a causal concept, it appears that using participation as the basis for responsibility might implicate all agents who play a part in the causal web leading to harm. This would entail casting the net of responsibility too far. I argue that on a plausible interpretation of participation, this problem can be avoided.

References

Bratman, M. (1997) "Responsibility and Planning" in *The Journal of Ethics, 1*(1), pp. 27-43.

Bratman, M. (2013) *Shared agency: a planning theory of acting together*. New York, NY: Oxford University Press.

Gilbert, M. (2000) *Sociality and responsibility: new essays in plural subject theory*. Lanham, Md.: Rowman & Littlefield Publishers.

Gilbert, M. (2006) "Who's to Blame? Collective Moral Responsibility and Its Implications for Group Members" in *Midwest Studies In Philosophy, 30*(1), pp. 94-114.

Gilbert, M. (2014) *Joint Commitment: How We Make the Social World*. New York, NY: Oxford University Press.

Isaacs, T. (2011) *Moral responsibility in collective contexts*: New York : Oxford University Press, 2011.

Kutz, C. (2000) *Complicity: ethics and law for a collective age*. Cambridge: Cambridge University Press.

Lawson, B. (2013) "Individual Complicity in Collective Wrongdoing" in *Ethical Theory & Moral Practice, 16*(2), pp. 227-243.

McKeown, M. C. (2015) *Responsibility without guilt: a Youngian approach to responsibility for global injustice.* (Electronic Thesis or Dissertation), UCL (University College London).

Petersson, B. (2013) "Co-responsibility and Causal Involvement" in *Philosophia, 41*(3), pp. 847-866.

Pettit, P. (2001) "Collective Intentionality" in R. J. O. a. J. W. Ngaire Naffine (Ed.), *Intention in Law and Philosophy* (pp. 241-254). Dartmouth: Ashgate.

Pettit, P. (2007) "Responsibility Incorporated" in *Ethics, 117*(2), pp. 171-201.

Young, I. M. (2011) *Responsibility for Justice*. New York: Oxford University Press.

Anneli Jefferson (University of Birmingham): Don't Look back in Anger – A defence of instrumentalist accounts of moral responsibility

When we think about responsibility, we always have one eye to the future and one eye to the past. We look at agents' past behaviour and intentions to establish desert, and we look to the consequences it is appropriate to visit on them. In his seminal 1968 paper 'Free will, praise and blame' Smart put forward an instrumentalist account of what it means to be morally responsible. In a nutshell, Smart's claim is that the ascription of moral responsibility as well as praise and blame are justified if they have the desired effect of improving the agent's behaviour. Smart concluded with satisfaction that on his account

‘the ascription and non-ascription of responsibility have therefore a clear pragmatic justification which is quite consistent with wholehearted metaphysical determinism” (Smart 1968, p. 302) Others, however, have been less impressed.

Probably the most common objection to Smart’s account is the following: Whether we take someone to be morally responsible and whether we should punish (or reward) their behaviour are separate questions, which can in principle be answered separately. For example, we can decide that somebody is responsible without at the same time blaming them, because we might think that certain forms of blame and punishment are counterproductive or problematic for certain reasons. (cf. Pickard 2011)

In other words, the problem with instrumentalist, forward-looking accounts of moral responsibility is that they do not distinguish between on the justificatory status of responsibility ascriptions and the usefulness (or lack thereof) of certain practices of holding responsible, punishing or rewarding.

In this paper, I take up and develop a proposal by Manuel Vargas according to whom instrumental considerations justify our practices of ascribing responsibility and of sanctioning immoral behaviour. I take his account to be the most promising (somewhat) instrumentalist account of responsibility available. After introducing the account, I discuss how well equipped it is to cope with the central objection to instrumentalism, that it does not distinguish between the warrant and the usefulness of blame and praise, reward and punishment.

The problem: While we frequently conflate judgments of moral responsibility with practices of punishment and reward, these are conceptually distinct and can come apart in specific cases. For example we make judgments regarding the moral responsibility of historical figures who we can no longer reward or punish (cf. Vargas 2008). Conversely, we sometimes treat someone as responsible, even when we are not convinced that they are fully responsible. Examples for this are the treatment of addicts (Charland 2011) but also our reaction to children’s moral and immoral behaviour. As we can see from these examples, our understanding of responsibility is not exhausted by considerations regarding the usefulness of holding responsible.

Manuel Vargas argues that the justification of our responsibility practices is instrumental. ‘Appropriately holding an agent responsible involves rightly regarding them as a responsible agent and correctly applying the justified norms of praise and blame, norms that derive their justification from their collective effects on responsible agency.’ (Vargas 2008, p.99)

Importantly, Vargas takes instrumental theories of moral responsibility to constitute only one building block in theories of moral responsibility. They specify the goal of our system of responsibility beliefs and practices (cf. Vargas 2008, Vargas 2013). According to Vargas, what responsibility practices ultimately aim at is to develop human moral responsibility, both that of the people attributing responsibility and of the objects of attribution. The further building blocks which are needed are 1.a theory of moral agency which specifies which capacities an agent needs to possess to count as morally responsible, a theory of the responsibility norms which specify when moral praise, blame etc. are justified. The role of instrumentalism is therefore more indirect than in Smart’s classic account, it is the general tendency to promote moral agency that determines our responsibility practices, not the effect in on one specific agent in one specific situation. While the justification for our responsibility practices is ultimately instrumentalist, it

takes a detour via human psychology and claims that the justification for our current practices of ascribing responsibility is that it is suited to make us better and more responsible people.
I show that this approach has the potential to incorporate important aspects from other approaches to moral responsibility, such as norm-expressivist theories and reactive attitude accounts. However, it does raise the question whether it can provide us with a non-consequentialist notion of when a responsibility ascription is justified. Vargas clearly states that it is not what we can hope to achieve by ascribing responsibility to an agent that underwrites ascriptions of responsibility. "On the account that I propose, whether the agent is morally responsible for his or her actions is not a function of a particular agent's susceptibility to influence in that particular circumstance, but rather a function of what the justified norms of moral influence say about the status of responsible agents in those contexts." (p. 103)
I argue that while a more indirect form of instrumentalism can indeed bypass many of the implausibilities of Smart's more crude instrumentalism, it cannot avoid the consequence that in as far as there are agents who are not susceptible to being influenced by our practices of holding responsible, it follows that these are not suitable targets for our responsibility ascriptions and practices. For example, if it turned out to be true that psychopaths are incapable of insight into the wrongness of their actions and immune to moral influence, they would not be morally responsible for their actions. However, the important (instrumental) role moral responsibility ascriptions play in developing and reaffirming our own moral stance also explains why this is so counterintuitive for many of us.

Mickelson, Kristin (University of Minnesota, Morris): The End of Incompatibilism

Standard arguments for free-will "incompatibilism"—e.g. the Consequence Argument (van Inwagen 1983) and the Manipulation Argument (Pereboom 2001; Mele 2006; Mickelson forthcoming)—conclude (at best) that *necessarily, if determinism is true, then the thesis that someone performs a free action is false*. As such, no standard argument for incompatibilism tells us what *makes it the case* that no one performs a free action in a universe at which determinism is true—*unless* the argument is supplemented with a best-explanation argument.

It might seem obvious that any such a best-explanation argument would have to pick out deterministic laws as (at least part of) the threat to free will. However, any plausible best-explanation argument that picks out deterministic laws as a specific threat to free will requires two highly controversial assumptions: (1) there is a viable libertarian-friendly analysis of *free will*, and (2) that some metaphysically possible being satisfies the necessary and sufficient conditions for free action specified in this libertarian-friendly analysis. Roughly put, only by arguing that indeterminism "helps" an agent to overcome the freedom-undermining feature(s) that obtain when determinism is true can one motivate the conclusion that deterministic laws are (part of) what "hurts". So, for those unwilling to grant assumptions (1) and (2), no standard argument for incompatibilism is plausibly understood as an argument for the view that (being subject to) deterministic poses a threat to free will.

Interestingly, though, if we do not grant (1) and (2), we can interpret every standard argument for incompatibilism as an incomplete argument for "constitutive-luck impossibilism", i.e. the view that free will is impossible because no possible agent can overcome the problem of constitutive luck. Once the goal shifts from explaining the mere *incompossibility* of free action and determinism to the *impossibility* of free action *tout court*, we cannot help ourselves to the assumption that it is metaphysically possible for someone to act with a (libertarian-style) free will. As such, one can no longer help themselves to the assumptions needed to generate a plausible best-explanation argument for the conclusion that deterministic laws play a metaphysical role in undermining free will. However, the constitutive-luck impossibilist can still explain the lack of free will in deterministic universes without arguing that deterministic laws pose a threat to free will. For instance, variants of the Basic Argument (e.g. Strawson 1998, Levy 2011) positively identify *constitutive luck* as an insurmountable threat to free will—and *without* reliance upon best-explanation reasoning. In fact, variants of the Basic Argument are the only extant arguments which illuminate *why* free action is impossible when determinism is true without relying on best-explanation reasoning or the assumption that free-will is metaphysically possible.

Notably, assuming that the constitutive-luck explanation is the correct, it discredits all theories which pinpoint less metaphysically fundamental features of the world (e.g. facts about its diachronic evolution) or its agents (e.g. whether they are subject to the laws or not) as playing a role in undermining free will (Mickelson 2015b: 2923-2925).[2] In other words, proponents of constitutive-luck impossibilism can hold that all extant arguments for incompatibilism—setting aside the supplementary (covert) best-explanation arguments and libertarian-friendly assumptions that are often used support them—support the same conclusion: the so-called "problem of determinism" is simply the problem of *constitutive luck*.

If I'm right, then all those (compatibilists, agnostics, and committed impossibilists) who are unwilling to grant the libertarian-friendly assumptions (1) and (2) should agree that all standard arguments for incompatibilism are best interpreted as (at least potentially) incomplete arguments for constitutive-luck impossibilism. As such, when incompatibilism is understood as the view that it is impossible to act freely in a universe at which determinism is true, *incompatibilism is a metaphysically arbitrary position*; when incompatibilism is understood as some version of the explanatory thesis that no one like us (i.e. beings subject to the laws of nature) could act freely in a universe with deterministic laws due to the deterministic laws, then compatibilists and impossibilists should agree that every major "argument for incompatibilism" concludes

[2] I argue that Neil Levy's (2011) puported "luck pincer", which identifies indeterministic causation as a threat to free will, rests on the assumption that counterpossible reasoning is coherent. In short, if constitutive luck makes free will impossible, then it would be metaphysically impossible for indeterministic causation to make someone unfree. In that case, assuming standard possible worlds semantics, it is false that it is possible for indeterministic causation ("present luck")to undermine free will.

that *incompatibilism is false*. To reject (1) and (2), then, is to reject incompatibilism as a metaphysically viable position.[3]

Put another way, I am proposing that all libertarian theories of free-will are best understood as proposed solutions to the problem of constitutive luck, and these solutions play an essential role in generating the idea that "deterministic laws" or "being subject to the laws of nature" are freedom-relevant properties (insofar as each purportedly *prevents agents from overcoming* the otherwise surmountable problem of constitutive luck).[4] In short, I contend that we can understand all the major arguments for incompatibilism as illuminating the same lesson: the "problem of determinism" is fundamentally the problem of constitutive luck; the only substantive disagreements are about whether and how the problem can be solved.

Finally, the arguments in this paper can be extended to shed new light on various outstanding problems in the free-will debate. For instance, my view that the Consequence Argument is best as understood as an incomplete argument for constitutive-luck impossibilism suggests a new solution to Joseph Campbell's "No-past Objection", for it would explain why standard arguments for incompatibilism are mysteriously weaker than commonly thought (Campbell 2007, 2008, 2010, 2011; Bailey 2013). My arguments could also be used to explain why the Consequence Argument and the Basic Argument seem to rely on a transfer *principle* (Campbell 2011) even though neither implies that deterministic laws are not a freedom-undermining transfer *mechanism* of non-freedom and non-responsibility (for constitutive luck alone provides adequate support for the relevant transfer principles). Finally, it suggests that the Manipulation Argument can be understood as a summation of the free-will debate, insofar as it can be used to explicitly explore the essential role that best-explanation arguments play in nearly all of the major moves and counter-moves in the history of the free-will debate.

References

Bailey, Andrew (2013). "Incompatibilism and the Past", *Philosophy and Phenomenological Research*
85.2:351-76.

Berto, F. (2013). "Impossible worlds", in E. N. Zalta (Ed.), The stanford encyclopedia of philosophy
(Winter 2013 Edition).
http://plato.stanford.edu/archives/win2013/entries/impossible-worlds/

---

[3] While the resulting view of the dialectic looks similar to the one advocated by Kadri Vihvelin (2013), I argue elsewhere that her arguments for this general view of the dialectic fail (Mickelson 2015a).

[4] I motivate the need to draw the distinction between the thesis of determinism and the thesis that the laws of nature are deterministic using an example of a "cosmic bruising" of a universe in a multiverse (Feeney, et all 2010). Once drawn, this distinction illuminates the importance of distinguishing between agents who are subject to the laws of nature and those who are not.

Campbell, Joseph (2007). "Free will and the necessity of the past", *Analysis* 67: 105–111.

—— (2008). "Reply to Brueckner", *Analysis* 68: 264–269.

—— (2010). "Incompatibilism and fatalism: reply to Loss", *Analysis* 70: 71-76.
—— (2011). *Free Will*. Massachusetts: Polity Press.

Feeney, Stephen *et al* (2010). "First Observational Tests of Eternal Inflation", http://arxiv.org/abs/1012.1995.

Levy, Neil (2011) *Hard Luck: How luck Undermines Free Will and Moral Responsibility*, New York: Oxford University Press.

Kane, Robert (1986) *Significance of Free Will*, Oxford University Press.

Lehrer, K. (1960). *Ifs, cans, and causes*. Dissertation, Brown University. Providence: ProQuest/UMI. (Publication No. AAT: 6205755.)

Loss, Roberto (2009). "Free will and the necessity of the present". *Analysis* 69: 63–69.

—— (2010). "Fatalism and the necessity of the present: reply to Campbell". *Analysis* 70: 76–78.

McKenna, M. (2010). "Whose argumentative burden, which incompatibilist arguments?—Getting the dialectic right". *Australasian Journal of Philosophy*, 88(3), 429–443.

-----(2012). "Moral responsibility, manipulation arguments, and history: Assessing the resilience of nonhistorical compatibilism", *Journal of Ethics*, 16(2), 145–174.

Mele, A. (2006). *Free will and luck*. New York: Oxford University Press.

-----(2008). "Manipulation, and moral responsibility", *Journal of Ethics*, 12(3), 263–286.

-----(2012). "Manipulation, moral responsibility, and bullet biting", presentation at the workshop on manipulation arguments, Central European University. http://humanproject.ceu.hu/events/2012-06-07/workshop-on-the-manipulation-argument

Mickelson, Kristin (forthcoming). "The Manipulation Argument", in *The Routledge Companion to Free Will*, eds. Meghan Griffith, Neil Levy, and Kevin Timpe.

-----(2015b) “The Zygote Argument is Invalid–Now What?,” *Philosophical Studies*. doi:10.1007/s11098-015-0449-6.

----- (2015a) “A critique of Vihvelin's Threefold Classification,” *Canadian Journal of Philosophy*, 45:1, 85-99, doi: 10.1080/00455091.2015.1009321.

Pereboom, Derk (2001). *Living Without Free Will*, Cambridge: Cambridge University Press.

Sartorio, Carolina (2015). “The Problem of Determinism and Free Will is Not the Problem of Determinism and Free Will”, in *Surrounding Free Will* Alfred Mele (ed.): 255-273.

Sehon, Scott (2010). “A Flawed Conception of Determinism in the Consequence Argument, *Analysis* 71.1: 30-38.

-----(2011). “What must a proof of incompatibilism prove? *Philosophical Studies* 154:361–371

Strawson, G. (1986). “Freedom and belief. Oxford: Clarendon Press.

-----(1994). “The impossibility of moral responsibility. *Philosophical Studies*, 75: 5–24.

-----(1998, 2011). “Free will”, in E. Craig (Ed.), *Routledge Encyclopedia of Philosophy*, London: Routledge. http://www.rep.routledge.com/article/V014SECT3

-----(2002). “The bounds of freedom. In Kane 2002 (pp. 441–460).

-----(2008). “The impossibility of ultimate moral responsibility,” in D. Pereboom (Ed.), *Free Will*, Indianapolis: Hackett Publishing Company: 289–306.

van Inwagen, Peter (1983). *An Essay on Free Will*. Oxford: Oxford University Press.

----- (2008). “How to Think About the Problem of Free Will”, *Journal of Ethics* 12(3/4): 327–341.

Vihvelin, K. (2013). *Causes, laws, and free will: Why determinism doesn’t matter.* New York: Oxford University Press.

Zagzebski, Linda 2002. “Recent Work on Divine Foreknowledge and Free Will”, in *The Oxford Handbook of Free Will*, ed. Robert Kane., Oxford: Oxford University Press: 45-64.

Zhang, Jiji 2013. Can the Incompatibilist Get Past the No Past Objection? *Dialectica* 67 (3):345-352.

Radoilska, Lubomira (University of Kent): Responsible Agency and Weakness of Will

In this paper, I revisit the debate between volitional and non-volitional conceptions of responsibility by focusing on agency in instances of weakness of will. I argue that a compelling account of responsibility for weak-willed actions points to a third, more fundamental model of action, action as actualisation, in addition to the models of action as production and action as assertion presupposed respectively by the volitional and non-volitional conceptions. The argument proceeds as follows.

In Section 1, I identify and explore some points of convergence between competing accounts of weakness of will (e.g. Mele 1987; Holton 2009): 1) responsible behaviour and 2) a form of criticisable irrationality. I reflect on the difficulty of capturing the irrationality of weakness of will without turning it into a concealed form of either rationality or non-rationality (Davidson 2001): weakness of will is still acting for a reason, though in a paradoxical kind of way. And so the target of criticism for weakness of will is difficult to pin down. Here are some possible contenders: constitution of agency, character flaw, unreliability, being less than fully responsive to reasons, loss of control. I propose a possible way forward: looking closer into alternative conceptions of responsibility. I then motivate two theoretical desiderata: 1) it is unfair for an agent to be held responsible for something that is not up to her and 2) attitudes are acknowledged as possible targets of criticism, on a par with actions.

In Section 2, I consider volitional conceptions, which interpret responsibility in terms of voluntary control. In particular, I discuss a reactive attitudes account put forward in Wallace (1994). According to this account, to understand responsibility, we need to understand first reactive attitudes, such as resentment and gratitude. I argue that this volitional account satisfies the first desideratum by deriving the conditions under which it is fair for an agent to be held responsible from a strong notion of reflective self-control. Yet, this is achieved at the expense of the second desideratum: attitudes are dismissed as unsuitable target for criticism, viz. responsibility ascriptions to the extent that they cannot be directly controlled by the conscious will. As a result, the volitional conception fails to provide a compelling account of responsibility for weakness of will. Instead, weakness of will is made to look like a concealed form of rationality: appreciating the better reason and choosing to act against it.

In Section 3, I consider non-volitional conceptions of responsibility, which replace voluntary control with evaluative judgment or quality of will. More specifically, I engage with the account presented in Smith (2005). According to this account, responsibility for

attitudes is a central case since attitudes and patterns of awareness more broadly can exhibit a rational connection to the person's evaluative judgments. I argue that this non-volitional account neatly satisfies the second desideratum; however, this is achieved at the expense of ignoring the first desideratum. As a result, the non-volitional conception also fails to provide a compelling account of responsibility for weakness of will. Instead, weakness of will is made to look like a concealed form of non-rationality, on a par with phobias, since there isn't a rational connection between behaviour and evaluative judgment.

In Section 4, I take stock of the discussion so far: both volitional and non-volitional approaches are meant to offer comprehensive accounts of responsibility. Yet, each seems able to do well in some central cases, but not others. Could there be a third, more fundamental conception of responsibility bringing together insights from these two apparently conflicting approaches? As a first step, I consider the models of acting responsibly implied by the alternative conceptions. I argue that while volitional accounts conceive acting responsibly in terms of production, non-volitional accounts conceive it in terms of assertion. On the first model, the point of acting is to bring about an effect: hence, the salience of voluntary control. On the second model, the point of action is to assert the agent's evaluative stance. Here, a link to this stance is more salient than control. If this is correct, the volitional and non-volitional accounts of responsibility turn out not to be mutually exclusive, since the models of action they build on are not mutually exclusive. Typically, responsible actions assert the agent's evaluative stance by bringing about some effect. Nevertheless, production and assertion are two aspects of success in action that sometimes come apart. I conclude by proposing a more fundamental model of acting responsibly: the actualisation model, e.g. by writing well, a person both asserts her positive evaluative stance with respect to the activity undertaken and ensures that the work she produces is good, while at the same time becoming or being a good writer. The actualisation model supports a compelling account of responsibility for weakness of will: weak-willed actions are successful as productions to the same extent that they are unsuccessful as assertions (cf. Radoilska 2013). At the same time, it points to a wider category of necessarily less than successful actions that cannot be fully successful, but are actions nevertheless. A major advantage of this model is that it captures well criticisable irrationality as distinctive feature of weakness of will. In so doing, it satisfies both theoretical desiderata identified earlier.

References

Davidson, D. (2001). Essays on Actions and Events. Oxford: Clarendon Press

Holton, R. (2009). Willing, Wanting, Waiting. Oxford: Oxford University Press

Mele, A.R. (1987). Irrationality. New York: Oxford University Press

Radoilska, L. (2013) Addiction and Weakness of Will. Oxford: Oxford University Press

Shapiro, T. (2001) 'Three Conceptions of Action in Moral Theory'. Noûs 35: 93-117

Smith, A. (2005) 'Responsibility for Attitudes: Activity and Passivity in Mental Life'.

Ethics 115: 236 -71

Wallace, R.J. (1994) Responsibility and the Moral Sentiments. Cambridge (Mass.): Harvard University Press

Robichaud, Philip (Delft University of Technology): The Threat of Choice Architecture to Moral Responsibility: Merely a Facade?

'Nudges' are cues inserted into a person's choice environment that lead to predictable effects on behavior. For example, employees might be nudged by open workspaces to increase their productivity and consumers might be nudged to purchase items whose high profit margin is linked to unethical production. Because nudges work by influencing automatic and unconscious modes of thinking, people are typically unaware that their choices are nudged. While many nudged actions are morally innocuous, others bring about significant harms or benefits. Much attention has been given to questions about the moral permissibility of nudging (Blumenthal-Barby and Burroughs 2012; Sunstein 2014). Of central concern are (i) the threats that nudging poses to authenticity and autonomy (Anderson 2010; Wilkinson 2013) and (ii) the compatibility of nudging with respect for human dignity (Waldron et al. 2014). A related but distinct question has received far less attention: do nudged agents deserve blame when they cause harm and praise when they realize some good? Importantly, all of the concerns that have been articulated as relevant for the permissibility question fail to speak directly to the question of whether agents deserve moral praise or blame for nudged actions. This omission is notable given that whatever answer ethicists come to concerning the moral permissibility of nudging, the responsibility question is arguably more urgent, given that engineers, marketers, and policy makers are already developing and employing nudges widely. In this paper, I take some steps toward exploring how certain nudges might impact moral responsibility attributions according to two influential compatibilist theories.

According to **control-based** theories of responsibility, agents are responsible for their actions only if the mechanism that causes the action is responsive to sufficient reasons to act. This entails that when there are strong reasons to perform a particular action, the agent must have the capacities to (a) recognize these reasons, and (b) bring herself to perform that action in a sufficiently broad range of circumstances. Call this the capacitarian condition. Control theorists also defend a historical condition according to which these capacities must be an integral aspect of the agent's identity and not the product of objectionable manipulation (Fischer and Ravizza 2000). So, in order to determine whether an agent deserves, say, approbation for a nudged accomplishment, it is important to understand clearly (i) how interventions on the decision-making environment affect the agent’s responsibility-relevant capacities and (ii) whether it can still be said that the capacities that are targeted by the nudge are integral to the agent. Because of the mechanism by which much nudging works – via unconscious and automatic forms of reasoning that are exploited by nudgers - both the capacitarian and historical conditions are plausibly under threat. In the first part of this paper, I focus my

attention on the capacitarian condition. I consider and reject an argument that nudged agents who act on the basis of unconscious processes or attitudes necessarily act from non-reasons-responsive mechanisms (Levy 2014a). Roughly, the argument holds that such agents fail to satisfy both components of the capacitarian condition. They fail to recognize the relevant reasons faced in the decisional context, and they would fail to perform the nudged action in the relevant subset of counterfactual worlds (i.e. those in which the nudge is absent). I challenge a central premise in this argument by drawing on the comparison between nudged actions and standard cases of rational deliberation in which the weighing of reasons and ranking of options occurs below the surface of awareness. I argue that agents for whom important steps in the deliberative process occur unconsciously meet the capacitarian constraint straightforwardly. I then show how the same considerations support the claim that certain nudged agents also meet it. If these arguments succeed, it follows that although it's true that nudged agents may lack conscious awareness of the features in virtue of which their actions are praiseworthy or blameworthy, they may nonetheless act from a mechanism that is reasons-responsive, and they may indeed be morally responsible according to control-based views.

In part two of the paper, I shift my attention to **quality of will** theories of responsibility, according to which agents are responsible for their actions only when they express some relevant set of desires or values that are taken to be constitutive of the agent's moral standpoint (for a recent flavor of this type of view see Arpaly and Schroeder 2013). On such views, it suffices for moral responsibility that the action expresses the agent's normative stance. In order to determine whether nudged agents are responsible on such views, one must test whether a nudged action brought about by the third-party exploitation of unconscious biases might nonetheless reflect her moral stance on things. In this section, I consider an argument that actions with unconscious moorings, such as those that result from implicit biases, fail to express the agent's evaluative stance (Levy 2014b). A central premise in the argument is that the unconscious attitude that is causally efficacious in some nudged action must interact to a sufficient degree with the relevant subset of the agent's attitudes. After considering some objections to this premise, I offer a refinement of it that leaves open the possibility that certain causally effective unconscious attitudes are indeed sufficiently integrated into the set of attitudes that make up the agent's moral stance. I go on to show how this position allows for the classification of some nudges as inimical to morally responsible agency (e.g. those that exploit framing effects) and others perfectly compatible with it (e.g. those that exploit status quo bias).

It turns out that the answer to the question posed in the title is: it depends.

References

Anderson, Joel. 2010. "Nudge: Improving Decisions About Health, Wealth, and Happiness, Richard H. Thaler and Cass R. Sunstein." *Economics and Philosophy* 26 (03): 369–76.

Arpaly, Nomy, and Timothy Schroeder. 2013. *In Praise of Desire*. Oxford University Press.

Blumenthal-Barby, J. S., and Hadley Burroughs. 2012. "Seeking Better Health Care Outcomes: The Ethics of Using the 'Nudge.'" *The American Journal of Bioethics* 12 (2): 1–10.

Fischer, J. M, and M. Ravizza. 2000. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.
Levy, Neil. 2014a. *Consciousness and Moral Responsibility*. Oxford University Press.
———. 2014b. "Consciousness, Implicit Attitudes and Moral Responsibility." *Noûs* 48 (1): 21–40.
Sunstein, Cass R. 2014. "The Ethics of Nudging." SSRN Scholarly Paper ID 2526341. Rochester, NY: Social Science Research Network. http://papers.ssrn.com/abstract=2526341.
Waldron, Jeremy, Samuel Freeman, Cass R. Sunstein, and Jerome Groopman. 2014. "It's All for Your Own Good." *The New York Review of Books*, October 9.
Wilkinson, T. M. 2013. "Nudging and Manipulation." *Political Studies* 61 (2): 341–55.

Schaab, Janis David (University of St Andrews/Stirling):

Commitment and the Second-Person Standpoint

Stephen Darwall suggests that the supreme principle of morality, which he identifies as Kant's Categorical Imperative, can be derived from necessary presuppositions of the second-person standpoint, i.e., "the perspective [we] take up when we make and acknowledge claims on one another's conduct and will" (2006: 3). He thus diverges from the strategies of other contemporary moral philosophers in the Kantian tradition, most notably Christine Korsgaard (1996, 2009), who think that the moral law can be inferred from the constraints of the standpoint of merely first-personal deliberation which all agents, *qua* agents, inevitably take up when they deliberate about what to do.

According to Darwall (2006: especially ch. 9), his strategy is superior to the one proposed by Korsgaard and others since practical reasoning from the second-person standpoint, unlike merely first-personal reasoning, necessarily presupposes that the reasoner possesses autonomy of the will, i.e., "the property of the will by which it is a law to itself (independently of any property of the objects of volition)" (Kant 1996: IV 440). The underlying idea is that the second-person standpoint is the locus of a specific type of reasons, second-personal reasons, whose normativity derives, not from the agent-neutral value of outcomes, but from relations of legitimate authority and accountability. That is, a second-personal reasoner acknowledges that she has reasons to act in certain ways precisely because others can legitimately demand that she do so, viz. hold her responsible for acting in these ways. According to Darwall, the reasoner thereby implicitly acknowledges that she possesses autonomy of the will and is thus bound by the Categorical Imperative. After all, to be guided by second-personal reasons an agent must be capable of freely determining her will by demands whose authority is independent of her antecedent desires or the agent-neutral value of states of affairs.

One potential problem for Darwall's view is that the second-person standpoint, unlike the standpoint of first-personal deliberation, does not seem to be inescapable for all agents (Darwall 2006: ch. 11, Hanisch 2014, Korsgaard 2007, Schapiro 2010). That is, an agent could in principle avoid ever taking up the second-person standpoint. It seems that, on Darwall's account, such an agent is not bound by the moral law. In this talk, I argue that, although the second-person standpoint and all its accompanying requirements are indeed escapable in principle (*pace* Darwall 2006: ch. 11, 2007, 2010), this is less problematic

than it might seem. In particular, I argue that it is even harder, psychologically speaking, to avoid taking up the second-person standpoint than has hitherto been noted. More specifically, I contend that we frequently take up this stance, not only in other-regarding, but also in purely self-regarding reasoning.

In order to show this, I appeal to Ruth Chang's work on commitment (2013a, 2013b). According to Chang, when we commit to a project, we employ our will's 'normative power' to give ourselves a reason to pursue the project which we would otherwise not have had. This reason is independent of the antecedent value or choice-worthiness of the project and therefore remains in place unless and until we withdraw our will's normative power by 'uncommitting' from the project. I contend that commitment thus exhibits the structure of second-personal reason-giving. Instead of deriving its normativity from the agent-neutral value of outcomes, it invokes a relation of authority and accountability—in this case, between ourselves and our own will. I therefore contend, *pace* Chang, that when we commit to a project we thereby take up the second-person standpoint vis-à-vis ourselves and thus implicitly acknowledge the authority of the moral law.

Although my account denies that the requirements of morality are inevitably binding on all agents, it establishes that the second-person standpoint is taken up in a larger range of cases than has hitherto been noted. In particular, it shows that we reason second-personally whenever we commit to a project, and thus not only in interpersonal, but also in intrapersonal contexts.

References

Chang, Ruth. 2013a. "Grounding Practical Normativity: Going Hybrid." *Philosophical Studies* 164: 163–187.

— 2013b. "Commitments, Reasons, and the Will." Russ Schafer-Landau (ed.), *Oxford Studies in Metaethics: Volume 8*. Oxford: Oxford University Press.

Darwall, Stephen. 2006. *The Second-Person Standpoint*. Cambridge, MA: Harvard University Press.

— 2007. "Reply to Korsgaard, Wallace, and Watson." *Ethics* 118(1):52–69.

— 2010. "Reply to Schapiro, Smith/Strabbing, and Yaffe." *Philosophy and Phenomenological Research* 81(1):253–64.

Hanisch, Christoph. 2014. "Self-Constitution and Other-Constitution." *Grazer Philosophische Studien* 90:105–30.

Kant, Immanuel. 1996. *The Cambridge Edition of the Works of Immanuel Kant: Practical Philosophy*, Mary J. Gregor (ed. and trans.). Cambridge, UK: Cambridge Universty Press. Page references are to page numbers of the Preußische Akademie Edition.

Korsgaard, Christine M. 1996. *The Sources of Normativity*. Cambridge, UK: Cambridge University Press.

— 2007. "Autonomy and the Second Person Wthin : A Commentary on Stephen Darwall's The Second-Person Standpoint." *Ethics* 118(1):8–23.

— 2009. *Self-Constitution: Agency, Identity, and Integrity*. Oxford: Oxford University Press.

Schapiro, Tamar. 2010. "Desires as Demands: How the Second-Personal Standpoint Might Be Internal to Reflective Agency." *Philosophy and Phenomenological Research* LXXXI (1): 229–36.

Stamets, George (Florida State University / University of Leeds)
"Free Choice, Control, and Acting for a Reason"

In contrast to event-causal and agent-causal views, non-causal varieties of libertarianism place no positive causal requirement(s) on free action; typically, they go further still in requiring that free actions – or at least 'basic' free actions, sometimes conceived as choices or volitions – be altogether *uncaused*. But few contemporary defenders of libertarianism accept a non-causal picture. Indeed, it is fairly common for non-causal libertarianism to be quickly cast aside as untenable at the outset of any conversation about the prospects for libertarian free will.

There are at least two major challenges facing non-causal libertarian accounts. First, there are a variety of 'luck' or control-based objections to libertarianism in general, according to which the indeterminism required by the libertarian introduces a degree of luck that rules out the special kind of *control* needed to say that a person has genuinely acted freely. While event-causal and agent-causal libertarians may try to answer this objection in part by appealing to indeterministic causation, this strategy is unavailable to the non-causalist. Second, there is the charge that non-causal libertarianism, since it typically eschews appeal to causation for the purpose of reason-explanation of intentional action, cannot accommodate the important notion of a person's acting – freely or otherwise – *for a reason*.

My purpose in this talk is to argue that the right sort of non-causal libertarian account – one that, among other things, embraces a powers theory of causation – can explain what it is to exercise agential control, and to act for a reason, better than any alternative account of free action, libertarian or otherwise.

Following E.J. Lowe, I sketch and motivate a view according to which ordinary human agents possess a *free power of will* – a literal and irreducible power to consciously and deliberately choose how to act, the exercises of which are necessarily spontaneous in the sense of being altogether uncaused. But while it is a spontaneous power and thus cannot be 'triggered' or 'stimulated', the will is not simply exercised, when it is, at random – for it is also what Lowe calls a *two-way*, *rational* power. A person's will is characteristically exercised by her in the light of various reasons for action of which she is aware, where such reasons are understood as external items – considerations that (at least apparently) justify, or speak in favor of, her acting in some way in a given set of circumstances. And the will – unlike any other power to be found anywhere in nature – is a 'two-way' power: it is a power to choose to φ or to refrain from φ-ing, for any action-type φ that a person should become aware of. This to say that a person's power to will to φ is the *very same* power as her power to will not to φ. Having such a two-way power gives one genuine alternative possibilities: consciously faced with the question whether to φ, a person can always, at a minimum, either choose to φ or choose not to φ. Which of these she performs, if any, will be an exercise of 'complete control': it will be that person's conscious and fully deliberate exercise of a spontaneous, two-way rational power, and nothing beyond her will play any role in causally influencing or determining that action.

Event-causal and agent-causal libertarians often lead the charge in expressing puzzlement as to how uncaused actions, like Lowe's volitions, could possibly constitute or involve exercises of the sort of control thought necessary for free action. Common to

such objections is the idea that a person's 'determining' some free action of hers must be a matter either of her causing it (qua substance-cause), or else of its being caused by certain events involving her (say, the onsets of relevant belief-desire pairs or intentions). But this widely-shared assumption is much too quick – and, indeed, is deeply problematic in its own right. Against standard agent-causal views in particular, I argue that the relation between a person and her free actions should be understood as internal rather than causal. Once we appreciate this point, we should recognize that a person's determining some volition of hers need not amount to its being caused by her, or else by certain events involving her, but may consist simply in her *spontaneously performing it*. It is this spontaneity of the power of will, together with its status as a two-way rational power, which entails that volitions are exercises of what I call complete control (and thus that volitions count as our basic free actions).

The second main challenge facing non-causal libertarianism is to explain how, if free actions are uncaused, a person can be said to freely φ for some particular reason R, rather than for some other reason of which she is aware (or perhaps for no reason at all). Most contemporary philosophers of action follow Davidson in adopting causalism about reason-explanation of intentional actions, according to which explanations that cite a person's reasons for acting intentionally must name the reasons that 'moved' (i.e., caused) the person to act as she did. I likewise reject this sort of causalism and, following Lowe, argue that we can make perfectly good sense of a person's choosing not only how to act, but on which reason so to act, without her being *caused* to do so.

Talbert, Matthew (West Virginia University):

Doing what you think is right

This paper considers the relationship between judgments of praiseworthiness and blameworthiness and the subjective perspectives of agents, particularly their views about whether certain actions are permissible. Some theorists believe that if a wrongdoer sincerely judges her behavior to be morally correct—and this judgment is rationally held because it is, for example, widely endorsed in her culture—then the wrongdoer has access to an excuse that insulates her from moral blame.[5] In a related fashion, an agent's belief that she has acted rightly is commonly cited as a form of moral praise, even if we believe that the agent in fact acted badly.

In this paper, I argue that an agent doing what she reasonably thinks is right is not, in itself, good grounds for excusing her from blame, nor does it do much to make her a fit target for praise. What matters for praise and blame, I argue, are not the subjective moral perspectives of the targets of these responses, but rather our moral perspective as issuers of praise and blame. On my view, the fact that an agent does what she thinks is right functions as a plausible excuse, or as plausible grounds for praise, only when the agent's moral views overlap with our own in some significant way.

The first section of the paper deals with an ambiguity in the idea that a person may

[5] I have in mind writers such as Susan Wolf, Neil Levy, Gideon Rosen, and Miranda Fricker.

wrongly believe that she acts permissibly. A person can be said to have such a belief if she is wrong about the consequences she anticipates from her action or, on the other hand, if she is wrong about the moral status of bringing about the consequences that she correctly anticipates. I argue that only the first interpretation of the proposal reasonably grounds excuse, and that it does so not because the agent's behavior conforms to *her* views about what is permissible or desirable but because her behavior conforms to *our views*—that is, the views of the blaming community. Suppose that someone rushes to the scene of an accident intent on helping an injured person but that, in her haste, she inadvertently makes the injuries worse. We might blame this person for acting rashly, but our blame may also be tempered by the thought that she did what she thought was right. On my view, this excuse is successful only to the degree that we take the actor to have been moved by sentiments that we endorse, such as a selfless desire to help the injured. The reference to the actor's moral views would be idle as an excuse except insofar as we find something of which we approve in the perspective that moved her to action.

There is much less basis, I argue, for excusing a wrongdoer who believes that she acts permissibly but who is moved by sentiments, or is trying to achieve ends, of which we strongly disapprove. This is because we have little choice but to interpret a wrongdoer's behavior as expressing contempt or ill will if she is moved by values or ends that we reject. The fact that the values by which the agent is moved are sincerely held and subjectively reasonable is irrelevant to our ascription of ill will because when it comes to making judgments about another's will, we have no choice but to rely on our own views about whether certain sentiments and goals are objectionable.

In the second section of the paper, I consider objections to the view just described. Some theorists say that it is unfair to blame wrongdoers who believe that they behave permissibly because it would be unreasonable to demand that these wrongdoers do otherwise. I agree about the unreasonableness of such a demand; however, I conclude that this points to a separation between what can be reasonably demanded of us and that for which we can reasonably be blamed. Since my view presupposes a tight fit between wrongdoing and blameworthiness, it may be thought to illegitimately collapse the distinction between being morally bad and being blameworthy. In addition, since I place heavy emphasis on the subjective moral perspective of those who issue blame, it may seem that I am committed to the unappealing claim that blame is reasonable if and only if it is judged reasonable by the one who issues it. I argue that these last two objections hinge on a misunderstanding of the view I present.

In the final section of the paper, I consider cases in which we praise others by noting their commitment to doing what they think is right—their integrity, in other words. This form of praise is related to the excuse discussed above, but they are not identical: the excusing force of the claim, "she thought it was the right thing to do," does not seem to derive simply from an ascription of integrity. Still, as with the putative excuse, the sort of praise envisioned here makes most sense when the goals and values of the one who is praised align substantially with those of the one who offers the praise. The reason for this is that, as the distance between our own values and those of the one we praise increases, her behavior must seem to have less to recommend it. In a case of extreme divergence between our values and those of the one we praise, the praise will be very faint indeed,

perhaps amounting just to the admission that the target agent does not possess certain defects, such as hypocrisy or inconstancy. For someone to be worthy of more substantial praise, we must see her as moved not just by her own values, but also by values that seem to us to be values worth having.

Tsai, George (University of Hawaii): The Role of Respect in Blame's Efficacy

Philosophical reflections on the relationship between blame's *efficacy* and its *justification* have focused mostly on whether blame can be justified on the basis of its efficacy—its desired-effect of modifying the attitudes and behavior of the blamed agent to comply with moral norms. For example, consequentialists have long defended an affirmative answer to this question (e.g., J.C.C. Smart[6]), while non-consequentialists have defended a negative answer (e.g., P.F. Strawson[7], T.M. Scanlon[8]). Rather than addressing whether blame's justification depends on its efficacy (a question I'm inclined to answer negatively - see footnote 4), this paper addresses how blame's efficacy may depend on its justification. Surprisingly, this latter question has been largely neglected. This despite the fact, obvious to keen observers of social interactions, that blame's efficacy depends on the blamed agent's perceiving or experiencing the blame as *justified*, as being supported by good reasons. Indeed, blame that is experienced by the blamed agent as *un*justified—that is, as unwarranted or inappropriate or unfair—typically generates resentment, resulting in responses and behavior in the blamed agent that run counter to or undermine blame's desired-effect.[9]

More generally, issues concerning the moral-psychological mechanisms underlying blame's aptness to produce certain attitudinal- and behavioral-modifying effects have been under-explored. These effects are important because they help to generate or sustain the motivationally shared reasons and norms that partly constitute valuable interpersonal relationships and moral and political communities. (That is, a sufficient degree of motivationally shared reasons and norms is required to realize a "thick" interpersonal relationship or social world.) To understand how blame functions in generating or sustaining the shared reasons and norms that partly constitute a relationship or community, then, we need to understand the moral-psychological mechanisms that enable blame to achieve its desired-effects.

---

[6] J.J.C. Smart, "Free Will, Praise and Blame." Hobbes and Hume both defend a version of blame as a kind of deterrent strategy—as justified if it is efficacious.

[7] P.F. Strawson, "Freedom and Resentment." Strawson is standardly interpreted as advocating a non-consequentialist approach to responsibility, but has also been interpreted as a consequentialist. For non-consequentialist readings, see Jonathan Bennett, "Accountability"; Gary Watson, "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme"; R. Jay Wallace, *Responsibility and the Moral Sentiments*; Stephen Darwall, *The Second-Person Standpoint: Morality, Respect, and Accountability*. For a consequentialist reading, see Victoria McGeer, "P.F. Strawson's Consequentialism."

[8] T.M. Scanlon, *Moral Dimensions*.

[9] This suggests that to justify blame on the basis of its efficacy is to sell it short. If blame is efficacious only on the condition that it is experienced as justified, then the justification of blame cannot be reduced to its efficacy. For the blamed agent who thinks blame is unjustified does not simply think that the blame will be ineffective.

Focusing on the central case of second-personal directed blame—understood as involving the outward communication of the reactive attitudes (resentment, indignation, moral anger) by the blaming agent to the blamed agent—this paper explores how blame is able to achieve its desired-effects in the blamed agent. In particular, I explore the relationship between: (1) blame's efficacy, (2) blame's justification, and (3) the attitude of respect between the blaming and blamed agents. I argue that blame's efficacy (its power to change the behavior of the blamed agent) depends in part on the blamed agent's willingness to accept certain justifications, and that this willingness on the part of the blamed agent may, in turn, depend on the blamed agent's possessing a broader, background desire to have or maintain the respect (or esteem or positive-regard) of the blaming agent.

My discussion draws on Bernard Williams's interpretation of blame as resting, in part, on a *fiction* that helps to recruit people into the moral community.[10] The fiction is that the agent who is blamed had reason to avoid the wrongdoing, comply with moral norms, at the very time at which the wrongdoing was committed. This is false, in William's view, given basic facts about the relation between (practical) reasons and the "subjective motivational sets" of the agents whose reasons they are.[11] But though the conception is a fiction, it can be a useful one that has a legitimate place in human social life. In particular, treating people *as if* this assumption were not false can be a useful and effective way of "recruiting" people to join the moral community; we treat them as if they were already members of the moral community all along, and this can give rise in them to the desires and concerns that will make them members of the moral community.

In developing Williams' view, I sketch the different but overlapping moral psychological mechanisms underlying blame's efficacy in both "standard" and "proleptic" cases of blame. In the more common "standard" case, the blamed agent actually had (sufficient) reason to avoid the wrongdoing at the time at which the wrongdoing was committed; the blamed agent had reason in virtue of having a relevant motivation in his or her "subjective motivational set." Blame's efficacy in these cases is explained by the fact that blame serves to *remind* the blamed agent of the existing reason and motivation. By contrast, in the special "proleptic" cases (which depend on blame, most of the time, working in the "standard" way), the agent who is blamed did not actually have reason to avoid the wrongdoing at the time at which she acted wrongly; this is because the agent did not have a relevant motivation in her "subjective motivational set" (no desire to avoid the particular wrong). So blame in this case cannot serve to *remind* the agent of the reason. Instead, it *creates* it.

But how, exactly, does blame in the "proleptic" case create reason? By *presupposing* it. That is, *in* blaming, the blaming agent presupposes that the blamed agent did have reason, and *by* so doing is able to create it for the blamed agent going forward. Proleptic blame's reason-creating power, I argue, is conditioned on the blamed agent's having a background desire to be respected by people whom, in turn, she

---

[10] Bernard Williams, "Internal Reasons and the Obscurity of Blame"; "How Free Does the Will Need to Be?" For helpful commentary, see John Skorupski, "Internal Reasons and the Scope of Blame."

[11] See Williams, "Internal and External Reasons"; "Internal Reasons and the Obscurity of Blame." Williams understands our acceptance of this falsehood as connected to our acceptance of a broader picture of ethical life in basic Kantian terms (the conception of the ethical he calls "morality"). See Williams, *Ethics and the Limits of Philosophy*, Ch. 10.

respects. (Simplifying things a bit, imagine our blamed agent has the broader motivation to avoid blame when it comes from people she respects, but otherwise no desire to avoid the particular wrong.) Given the broader motivation, combined with the recognition of what is expected of her by someone she respects (which the blame serves to indicate), the blamed agent comes to acquire the desire to avoid doing the particular wrong (the thing she is blamed for having done). Having acquired this relevant item in her "subjective motivational set," then, the blamed agent now has reason to avoid the particular wrong going forward.[12] By thinking through blame's operation in both "standard" and "proleptic" cases—and exploring blame's efficacy, more generally—my discussion illuminates how blame, as a mode of influence, operates in a space between coercion and deliberative-advice.

Wallace, Robert H. (University of Arizona):

"Freedom and Resentment" and Persons

Freedom and Resentment" has become a modern classic. It singlehandedly drew the attention of the philosophical community to features of moral responsibility that had been tragically unnoticed. For instance, one would be hard pressed to find a contemporary article on moral responsibility that did not address the moral emotions. Yet one of the most important arguments within that article, what I will call (following McKenna and Pereboom) the *argument from exculpation*, is no longer widely discussed. In this paper, I revisit the argument in order to defend it from criticisms that have made it seem utterly implausible, and to provide an alternative way of understanding the importance of Strawson's framework by drawing a link between his views about what it is to be a person and his views on agency and responsibility.

The argument as stated in "Freedom and Resentment" is simple: the truth of determinism does not provide grounds for excuses, justifications, or exemptions from moral responsibility. How can he advance such a simple argument? Firstly, Strawson believes that the practices of holding someone moral responsible is primarily affective in nature—i.e., the praising and blaming is expressed by, or consists of, emotional *reactive attitudes*. Secondly, Strawson holds a *quality of will* thesis, that being morally responsible is most fundamentally about the good or ill will with which a person acts, and so, when holding another responsible by way of a reactive emotion like resentment, we reacting to that good or ill will. Third, Strawson appears to support the following grounding thesis: the disposition to or practices of holding a person responsible grounds the conditions for their being responsible. At a

minimum, Strawson appears to think that our practices of holding one another responsible are explanatorily, if not metaphysically, prior to and more basic than our actually being responsible. With this framework in hand, Strawson discusses two different kinds of pleas in our moral responsibility practices. The first kind,

[12] The hard cases, those who are *beyond the reach of blame*, might then be understood as those who do not care about having the respect or esteem or favorable regard of relevant others.

encompassing excuses and justifications, attempts to show that a purportedly blameworthy agent did not act from ill will, or without a sufficient degree of good will. The second kind, exemptions, attempts to show that it is a mistake to think that the purportedly blameworthy agent is capable of having or expressing a morally salient quality of will. Thus, Strawson thinks, determinism cannot count as a universal plea to excuse us from, justify certain actions, or exempt us altogether from moral responsibility, so long as one accepts the anti-reductionist naturalism he proposes.

The argument as it stands faces two important challenges. Firstly, it seems that there are existing pleas in our moral responsibility practices that might count universally: "being unable to do otherwise" and "being unfortunate in formative circumstances," for instance. Secondly, Strawson did not fully specify the relevant kind of *incapacitation* that might make someone unable to have or express a morally relevant quality of will. Although some Strawsonian philosophers have attempted to defend Strawson against these charges, I suggest that these defenses are insufficient. I then move to a more robust defense of the argument.

I begin with Michael McKenna's contention that understanding others' actions as bearing meaning, as being the expression of inner attitudes, requires that we understand ourselves as acting in a way that expresses meaning. Then, I bring McKenna's conception of morally responsible agency as involving self and other-understanding together with Strawson's work on personal identity in *Individuals*. Strawson argues that understanding oneself as a person—a being for whom both bodily and mental predicates are ascribable,

involves viewing others as persons, and vice versa. I then argue that there is no single capacity that makes one a responsible agent, but rather, that a plurality of cognitive and affective powers, which rise and fall together, supports our ability to have and express quality of will—and be fully developed persons. Thus, I contend that the best way to be a Strawsonian naturalist is not to stop at the thought that our practices ground the conditions of morality responsibly, but rather, to think about those practices as themselves grounded in more basic natural facts about what it is to be a person.

By developing Strawson's position in the way I have suggested—by arguing that Strawson's own view of what makes one a moral agent involves a somewhat untold story about cognitive capacities in addition to the explicit one involving practices and reactive attitudes—I am moving away from the standard interpretation of Strawson. As such, I owe an account of how a contemporary Strawsonian might develop what I contend to be Strawson's full view. In the context of Strawson's *Skepticism and Naturalism*, I suggest that the overlooked cognitive capacities can be fleshed out in conjunction with contemporary views in cognitive psychology. In particular, we can articulate naturalistic theories about our ability to self-regulate and to understand the reactive attitudes. This discussion of psychology helps to resolve the other problem with the argument from exculpation related to existing pleas that seem to count universally, such as "being unfortunate in formative circumstances" and "being able to do otherwise." I argue that what the Strawsonian should be concerned with is any history that prevents persons from developing a *full* set of psychological know-how and affective susceptibility, and that a natural understanding of being able to do otherwise involves ordinary reactivity to other

persons. This line of thought leads to the following: in extreme situations, we should help others develop the capacities that make someone a

person, in the sense described above. Thus, I have hoped to tackle the two problems facing the argument from exculpation about capacities and universal pleas.

Given my reading of Strawson, his anti-reductionist naturalism turns out to be a position about the irreducibility of being a person. I close by considering the attractive features of such a view and the prospects for such a theory in the context of Strawson's own apparent relativism about the relationship between our ordinary way of seeing the world and an "objective" view of things.