Recasting the Problem of Resultant Luck

Marcelo Ferrante
Universidad Torcuato Di Tella
School of Law
Miñones 2159 (C1428ATG)
Buenos Aires, Argentina
mferrant@utdt.edu

February, 2008

Abstract

I offer in this paper an argument in support of the orthodox view that resultant luck should not affect judgments of blameworthiness—and so, e.g., that we shouldn't blame the successful assassin more than the attempted assassin who equally tries but fails. This view, though widely held among moral philosophers and legal scholars, has been severely challenged as implying either the implausible rejection of moral luck, or an equally implausible theory of wrongness according to which actual consequences may play no wrong-making role. The argument I offer, however, assumes both challenges to be true and shows that the orthodox view is consistent with holding them. Indeed I argue that all other things being equal successful offenders are no more to blame than their unsuccessful counterparts even though agents are responsible for what they actually do (and therefore are subject to moral luck), and successful offenders do more wrong than their unsuccessful counterparts do (and therefore consequences do play a wrong-making role). The reason is that the difference in the amount of wrong done by one and the other offender, I show, is outweighed by a difference in the degree to which the successful offense and the unsuccessful one are attributable to their respective agents—blameworthiness being a function of both amount of wrong done and degree of attributability.

Introduction

Both in morality and in criminal law, we condemn the successful assassin more intensely than we condemn the attempted assassin who equally tries but fails. The orthodox view among legal theorists and moral philosophers is that we do wrong: the unsuccessful offender is as blameworthy as the successful counterpart is. The "luck in the way actions turn out,"[1] or resultant luck, is of no significance to the issue of deserved blame. Insofar as our condemnatory practices are to track judgments of moral responsibility, we should make no distinction between ceteris paribus successful and unsuccessful offenders. Arguments of two different sorts have been advanced in support of this orthodox view. According to the first kind of argument, the orthodox view is an implication of the denial of moral luck—or, more narrowly, of the principle that we mistreat a person when we let factors she did not control influence whether, or how much, we condemn her. Arguments of the second sort show the orthodox view to be a reflection of a theory of wrongness under which there is no more wrong done in succeeding than in equally trying but failing. That's the case if, say, it is risking or intending harm, rather than actually harming, what wrongness depends upon.

Both the denial of moral luck and the equal wrongness claim, however, have proven far from convincing. The denial of moral luck has implications that the we are normally unwilling to hold. If the reason why we should condemn the pair of successful and unsuccessful assassins alike is that the success-failure distinction tracks a feature of the

[1] The expression is Nagel's. See Thomas Nagel, "Moral Luck," reprinted in *Mortal Questions* (New York: Cambridge University Press, 1979).

world that lies beyond the agents' control, and that such features cannot ground differences in blameworthiness, then we would be committed to equally condemn many others who did not actually try to murder, or even think about it. Indeed, the two assassins could not be more to blame than other individuals—almost everyone perhaps—who *would* have tried to murder just as the pair of assassins actually did try but for some feature of the world they did not control, like the lack of a propitious opportunity, the formation of a inconsistent intention, or the lack of the required character traits. If the denial of moral luck is what grounds the orthodox view, then it will entail a set of equally blameworthy people that is far larger than the orthodox is willing to accept—far too large indeed.[2]

One may try to escape this implication by basing the orthodox view on the equal wrongness claim. David McCarthy has argued that, for agents with our characteristic epistemic limitations, the concern with harm to others and other evil consequences of our acts should lead to a theory of wrongness according to which it is acts' *expected* consequences (or the *risks* acts create) the relevant wrong-making feature, rather than (some of) the consequences acts *actually* bring about.[3] It is hard to disagree with the idea, on which the argument depends, that for a feature of the world to count as a wrong-maker it must be somehow epistemically available.[4] However, it doesn't follow from it that the

[2] See Michael S. Moore, *Placing Blame* (Oxford: Oxford University Press, 1997), 191-247; Michael J. Zimmerman, "Taking Luck Seriously," *Journal of Philosophy* 99 (2002), 553.

[3] David McCarthy, "Actions, Beliefs, and Consequences," *Philosophical Studies* 90 (1998), 57.

[4] Id., 74.

relevant wrong-maker must be risk imposition, rather than actual causation of something like foreseeable consequences, or some suitable combination of both. If intending to kill my colleague next door, I throw a grenade into his office thereby killing him, it is not at all clear why we should give up the intuition that what makes my action wrong is (in part) that I kill my colleague as I intended, rather than (just) that I imposed risk of death on him.[5]

I offer in this essay a defense of the orthodox view that doesn't fall prey to the objections raised against the denial of moral luck and the equal wrongness claim. As I will argue, there is room for the orthodox view even if, plausibly enough, (i) people are blameworthy for what they actually do, which implies that they are subject to moral luck; (ii) blameworthiness, and the degree to which one is to blame depend partly on the wrongness of that which the agent is to blame for; and (iii) wrongness depends on external features, rather than only on internal features like beliefs and other intentional states, such that successful offenders may be said to do wrong in a way, or to a degree, that is more serious, or higher, than the wrong the unsuccessful counterpart does.

Success-Failure pairs

Arguments for the orthodox view normally involve a pair of cases of a characteristic sort, and mine will be no exception. I call these pairs *Success-Failure pairs*. Success-Failure

[5] I think that Thomson's objections in her "Imposing Risks" article still hold after McCarthy's argument. See Judith Jarvis Thomson, *Rights, Restitution and Risks* (W. Parent ed., Cambridge Mass.: Harvard University Press, 1986), 173-191.

pairs are supposed to make it salient that in treating successful and unsuccessful offenders unequally we afford significance to a feature that, we should recognize, has no significance at all. The first member of the pair, the *Success* case, describes the action of an agent, Alvarez, for which Alvarez is indisputably to blame. It might be that Alvarez kills her neighbor by shooting at her from the attic, or burns down her enemy's house by throwing an incendiary bomb—the details of the case are unimportant; you may bring in those of your choice. For convenience, I will use the variable "α" for the type of action that best describes Alvarez's doing in this context. The only crucial feature to observe in devising the *Success* case is that there must be some room for failure in the agent's α-ing. The α-ing must ultimately depend on conditions that are, as it were, "up to nature."[6] Human actions in the actual world are always like that, so any realistic case will do. The reason for this condition is that the *Failure* case, the second case of the pair, is to fall within the space of Alvarez's likely failure—the *Failure* case actualizes a possible but unactualized failure of the *Success* case. The agent in the *Failure* case, whom I will call Borges, is identical to Alvarez in every respect, displays the same skills, beliefs and intentions as Alvarez does, engages in an identical course of action, which is as likely to succeed as his counterpart's is, but, unlike Alvarez, Borges fails. What are the particular conditions of α-ing that the agents neglect is again unimportant: the neighbor's window might happen to have bullet-proof glass, or an unexpected passing bird might deviate the

[6] I play here with a sentence of Davidson's with which he characterizes basic or primitive actions: "We never do more than move our bodies: the rest is up to nature." Donald Davidson, "Agency," in *Essays on Actions & Events* (Oxford: Clarendon Press, 1980), 59.

shot; a sudden rainstorm might kill the fire, or a team of firefighters might happen to be partying in the house next door, with the fire truck and everything. The pair of cases must be such, in other words, that Borges does *everything* Alvarez does except α-ing, and her failure cannot be explained as the consequence of something one agent has done, or omitted to do, while the other has not. To simplify, I shall use the variable "β" for the type of action that best describes Borges's action. Both Alvarez *and* Borges thus do β, only that Alvarez's β-ing becomes an α-ing, while Borges's does not, in virtue of some non-actional difference between the cases.

Let me clarify a bit more what the relationship is between the types of actions (α and β) for which the agents (Alvarez and Borges) are to blame. First, β-ing is for the agents the *actional means* to α-ing, which entails the following two things: (i) β-ing is *more basic* an action than α-ing is (in the sense of "basic" in which we say, non-comparatively, that a given action is a "basic action"),[7] which means that for the agent to do α she must do β as *part* of it, or that she can only do α *by* doing β. (ii) β-ing is *practically sufficient* for α-ing, which is to say that the agent does α *just* by doing β. Not that β-ing *entails* α-ing—it does not; it is indeed practical rather than logical sufficiency that is at stake.[8]

[7] An action ϕ is a basic action of an agent S if and only if there is no other action φ such that S does ϕ by doing φ as part of doing ϕ. On the concept of basic action see Arthur C. Danto, "Basic Actions and Basic Concepts," reprinted in his *The Body/Body Problem* (Berkeley: University of California Press, 1999), 45-62.

[8] The β-ing may or may not be an *attempt* to α-ing as the term is used in legal parlance. Yet, if it is—for instance, if α is coextensive with a crime definition, and the agent acts with the required intention—it will

Second, the α label stands for the best description of Alvarez's behavior for which she is to blame. That is, the description picks out *every* feature of the agent's doing *for which* the agent is to blame—which, in turn, are the same as the features in virtue of which the doing is morally wrong. Particularly, when wrongness is to do with doing harm, wrongness turns not only on the fact that harm is caused, but also on the way the harm is produced. So, the α label is meant to include not just the fact that a harm of a particular kind has been produced some way or another, but also every aspect of the way in which the harm is brought about affecting the overall judgment of wrongness. Suppose I intentionally kill Martha, my colleague next door, by throwing a grenade into her office. I do wrong, no doubt, in virtue of the fact that I kill her; and assuming I have no excuse, I am to blame for killing her. Killing Martha is certainly the dominant wrong making factor in this story, but it is not the only one. There are also other features of my behavior which are intuitively relevant for an exhaustive judgment of wrongness. I think most people would agree on the following: that it is an intentional killing (rather than, say, an accidental killing), by a close colleague while the victim is working at the university (rather than, say, by a war enemy on the battlefield), and that it brings about death in a physically destructive way (rather than in a focused, painless, clean and quick way). If I

---

satisfy what criminal lawyers sometimes call "*complete* attempt." See e.g. Joshua Dressler, *Understanding Criminal Law* (2d ed., New York: Matthew Bender, 1995), 347-8. Hence, the orthodox view is *not* the view that criminal attempts should be punished as severely as the corresponding accomplished crimes. The orthodox view only warrants that, all other things being equal, the perpetrator of an accomplished crime is no more to blame than the perpetrator of the corresponding *complete* attempt.

am right, then, to blame me just for killing Martha would understate the ground of my blameworthiness. A more accurate description of that for which I am to blame would be something like "intentionally killing Martha, the colleague next door, by throwing a grenade into her office." Thus, if "I intentionally kill my colleague next door by throwing a grenade into her office" picks out every wrong making aspect of my action, then "I α" would stand for it.

The features of the action that the β description picks out, in turn, are determined by the α description and the notion of actional means. Thus, in the example of my killing Martha, where "I α" equals "I intentionally kill my colleague next door by throwing a grenade into her office," "I β" would stand for something like "I throw a grenade into my colleague's office with the intention to kill her." For I do α just by doing that.

I mean the proposition "Alvarez does α" to entail the proposition "Alvarez does β"—though not the converse—so that when we place blame on Alvarez for her doing α we are expressly reacting to the features of the case that make it true the β description (as well as the α description).[9]

[9] This point rules out the apparent oddity, pointed to my by Seana Shiffrin and Eduardo Rivera-López, in the orthodox claim that Alvarez is no more to blame than Borges is, when Alvarez would be to blame *both* for α-ing *and* for β-ing, whereas Borges is to blame only for β-ing. My point secures that although we can describe Alvarez's doing as an instance of α and an instance of β, we cannot place blame on Alvarez *both* for α-ing *and* for β-ing. Alvarez is to blame for α-ing. That is the relevant description under which she is to blame. Any other description would understate the grounds of her blameworthiness. We can re-describe her behavior more narrowly by saying that she does β. Under *that* description Alvarez would still be

Success-Failure pairs are expected to highlight the role of luck in the agents' doings. Alvarez succeeds, but he could as well have ended up as the agent of the *Failure* case. Likewise, Borges fails, but he could as well have starred in the *Success* case. Neither of them did anything that the other did not to guarantee one particular role rather than the other. Their actions turn out the way they do (α-ing or just β-ing) in virtue of sheer luck.[10] In other words, Alvarez's β-ing *happens* to turn out an α-ing, in the sense that it could as well have remained a *mere* β-ing, as it is the case in Borges's case. In *Success* the world is such that the β-ing becomes an α-ing. In *Failure* the world is minimally different, so that the β-ing fails to become an α-ing.

Alvarez cannot be more to blame than Borges is, the orthodox claims, because such differences in the agents' luck as to how their actions turn out cannot by themselves ground differences in blameworthiness. It is in the nature of blameworthiness that they cannot.

---

blameworthy (and no less blameworthy, according to the orthodox view), just as Borges is. But this is not to say that Alvarez may be blamed for two wrongdoings. Similarly, if I intentionally kill Martha by throwing a grenade into her office, I also kill Martha, period. My behavior makes me blameworthy both under the description "intentionally killing Martha by throwing a grenade into her office" and under the narrower description "killing Martha." Placing blame on me both for killing Martha and for intentionally killing Martha by throwing a grenade into her office would be to engage in a kind of double-counting.

[10] The impact of luck is higher if, say, the agent shoots playing Russian roulette than if she shoots after checking the gun to be fully charged. The luckier Alvarez is, the less extraordinary the corresponding Borges's case will be. Depending on the details of the case, Alvarez could be more or less lucky—but lucky she always is.

Blameworthiness, wrongness, and ownership

If blameworthiness turns on wrongness, then, all other things being equal, the more wrong an action is, the more blameworthy its agent will be. If, as I will be assuming, in Success-Failure pairs Alvarez's conduct is more wrong than Borges's conduct is, then, all other things being equal, Alvarez must be more to blame for her α-ing than Borges is for her β-ing. However, as I will argue here, things are not equal.

Blameworthiness is a function of *responsible* wrongdoing, not just wrongdoing—one can do wrong without being to blame for it, as when wrong is done under insanity or other excusing conditions. Responsibility for a given action, as well as wrongness, comes in degrees. So that one can be more or less to blame for a given wrongdoing depending on the degree of one's responsibility for that wrongdoing. Likewise, and this is the key to my argument for the orthodox view, one may be equally to blame for wrongs of different values, if the corresponding degrees of responsibility are such that they outweigh the difference in wrongness. Indeed, my argument will be that, although Alvarez does more wrong by α-ing than Borges does by β-ing, Alvarez's responsibility for α-ing is proportionally lower than Borges's responsibility for β-ing.

Being responsible here, as in the Strawsonian account of responsibility, is being appropriately susceptible to any of the reactive moral attitudes (e.g., love, admiration, grati-

tude, contempt, resentment, indignation, etc.).[11] Blameworthiness is a species of this more general notion of moral responsibility; being blameworthy is thus being appropriately susceptible to a particular reactive attitude—namely, blame. In this general sense, responsibility is, as David Copp put it, *response-worthiness*:[12] For an agent S to be responsible for an action ϕ of hers is to be deserving of a moral response (any one) in virtue of her ϕ-ing—that is to say, her ϕ-ing makes it fitting that S is reacted to with some of the characteristic reactive moral attitudes.

In this sense of response-worthiness, responsibility centrally depends on what we may call *attributability* or *ownership*: S is responsible for her ϕ-ing if and to the extent that the ϕ-ing is *S's own* in a strong sense.[13] Attributability only warrants that some moral response is due. For a particular moral response to be warranted further conditions need be met. Blameworthiness requires that that for which one is to blame makes it fitting the

---

[11] Peter Strawson, "Freedom and Resentment," reprinted in *Free Will* (G. Watson ed., Oxford: Oxford University Press, 1982), 59; John Martin Fischer, "Recent Work on Moral Responsibility," *Ethics* 110 (1999), 93-5.

[12] David Copp, "Defending the Principle of Alternate Possibilities: Blameworthiness and Moral Responsibility," *Noûs* 31 (1997), 441, 452.

[13] See Gary Watson "Two Faces of Responsibility," reprinted in Gary Watson, *Agency and Answerability* (Oxford: Clarendon Press, 2004), 260. Watson's argument is that attributability corresponds to a kind, or a face, of responsibility, the other kind (or face) being accountability. Other philosophers understand responsibility just in terms of attributability—see Fischer, n.11, 96, citing Derk Pereboom as an example. Whether attributability exhausts the conditions of response-worthiness or not I need not adjudicate here. I content myself with the view that attributability is at least part of the conditions of responsibility.

distinctively negative valence of the blaming reaction. The natural view is that for S to deserve blame for her ϕ-ing it must be that she ought not to ϕ, that it is morally wrong for S to ϕ. Thus, attributability accounts for the fact that S *deserves* blame for her ϕ-ing; wrongness, in turn, accommodates that it is *blame* what she deserves.

I follow Robert Nozick in framing the interplay of ownership and wrongness in terms of the product of a variable representing a measure of the seriousness of the action's wrongness times a discounting coefficient representing the degree of responsibility of the agent for the wrongdoing (ranging between zero, for no responsibility, to one, for full responsibility). Under this framework, an agent's degree of blameworthiness for an action ϕ of hers is a function of the product r_ϕ x W_ϕ, where W_ϕ stands for the measure of ϕ's wrongness, and r_ϕ is the discounting coefficient representing the degree to which the ϕ-ing is the agent's own.[14]

I contend that there is between the cases in Success-Failure pairs a difference in responsibility—one, moreover, that outweighs the difference in the amount of wrong done by one and the other agent. To show this I will analyze attributability in terms of responsiveness to reasons. Reasons-responsiveness, as exhibited in S's ϕ-ing, has two dimensions, I argue: one negative (*reactivity*), which involves the capacity to not-ϕ in response to reasons to not-ϕ; and the other positive (*reliability*), which involves the capacity to ϕ in response to reasons to ϕ. In Success-Failure pairs the reactivity variables are

[14] Nozick's framework was introduced in Robert Nozick, *Anarchy, State, and Utopia* (Basic Books, 1974), 59-63, and was later elaborated upon in Robert Nozick, *Philosophical Explanations* (Cambridge, Mass.: Harvard University Press, 1981), chapter 4.III.

fixed, as it were, at constant values. Reliability, in contrast, changes its value from one case to the other, and this variation accounts for the difference in attributability on which I base my argument for the orthodox view.

A simpler example may serve to illustrate the kind of difference in attributability that I argue obtains between cases in Success-Failure pairs. Suppose that in my first day at doing archery, I try with a first shot and my arrow happens to hit the bull's eye. And suppose that my archery instructor also shoots and hits the bull's eye—as she normally does when shooting from that distance. There is an intuitive sense in which the instructor may claim more responsibility for her hitting the bull's eye than I can claim for my own. The reason is that she has been far more in control of her shot than I have been of my own, and therefore there is more of hers in her hitting the bull's eye than there is of mine in my hitting the bull's eye. Her hitting the bull's eye is more her own doing than mine is my own doing. For luck has played a larger role in my hitting the bull's eye than in the instructor's, and the more an action owes to luck, the less it owes to the agent.

The archery examples are simpler than Success-Failure pairs in that we compare in them attributability for actions of an identical type—hitting the bull's eye—whereas in Success-Failure pairs we compare responsibility for actions of different types (α and β), which makes the difference less salient.

In the following two sections I build my case for this part of my argument for the orthodox view. Then I complete the argument by showing that the difference in responsibility outweighs the difference in wrongness between Alvarez's α-ing and Borges's β-ing.

On ownership

Attributability or ownership is a relation that links agents and actions under descriptions.[15] For one and the same act—understood as a particular event— there may be many things we may predicate of it as things that an agent does.[16] Yesterday night, for instance, I telephoned my old friend Gabriel—it was his birthday and I wanted to surprise him with a call after so many years. My call woke Gabriel's baby boy up, just after my friend had struggled for hours to sleep him. So, it is true of my action that it was a "telephoning Gabriel on his birthday" and a "disturbing the baby's sleep." Responsibility varies depending on the description under which the action is evaluated. Described as telephoning Gabriel on his birthday my action was something nice for me to do, something that called for a positive moral reaction—some sort of gratitude for my remembering him on his birthday. But it is also true of the very same event that it was a very late call, especially taking into account that Gabriel's baby boy was already asleep and that the ring woke him up. Gabriel was therefore resentful at my calling him so late, and he was right. For

[15] Not that attributability may not link agents with other items. All that I will say in regard to the attributability of actions may be applied, duly adjusted, to other things we may claim responsibility for, like beliefs or emotions. I restrict my focus to responsibility for actions just because my interest in this essay is to elucidate the issue of the relative blameworthiness of agents in Success-Failure pairs who only differ from each other in what they do.

[16] I will be assuming a coarse-grained view of the individuation of actions. Under a fine-grained view of the individuation of actions this point will not be in order. Nothing of importance turns on the choice of the act-individuation approach.

under the description "telephoning Gabriel that late thereby disturbing the sleep of his baby" my action was a poor thing for me to do, calling for therefore some kind of negative response. Resentful although he was for my calling him so late, Gabriel still thanked me for my calling on his birthday—and he was honestly thankful as far as I could tell. He wasn't irrational in thanking me. The positive judgment that Gabriel expressed in his thankfulness is fixed to whatever it is that makes it true of my action that it was a "calling Gabriel on his birthday." The negative judgment that he also expressed in his manifest resentment attaches to other aspects of my action, namely, those that make it true the description of it as an inadequately late call.

As responsibility varies with action-description, so does ownership. The attribution of an action ϕ, where the label "ϕ" stands for an act description, to an agent S depends on its being the case that (i) there is an event *e* such that *e* is an action of S's; (ii) *e* is ϕ—that is, we may predicate with truth "ϕ" of *e* or, what amounts to the same thing, we may speak with truth of *e* as an instance of ϕ or a ϕ-ing; and (iii) it is not just an accident for S that *e* is ϕ, that is, S exerts some suitable degree of control over *e*'s being ϕ. Conditions (i) and (ii) just secure that it is the case that S did actually ϕ.[17] It is the condition (iii), the control condition, that calls for some elaboration.

[17] Throughout I will refer to an instance of ϕ-ing that meets the first two conditions as an action of S's, even though it fails to meet the third condition—and so it is not S's own in the strong sense that would warrant a moral response. I don't mean to make any point on the use of the English language. I'm just trying to isolate, and then focus on, the control condition in any action for which an agent may be responsible. If what remains after we thus subtract the control condition is something we can still *properly* call an action, is a

There are two kinds of control philosophers distinguish as conditions of responsibility.[18] On the one hand there is what it is often called *alternative-possibilities control*. Alternative-possibilities control emphasizes that the ϕ-ing that actually obtains was one among other possible actions actually open to the agent. Roughly, S has alternative-possibilities control over her ϕ-ing, only if she could do other than ϕ-ing. On the other hand, there is what Timothy O'Connor has called *agent control*.[19] Agent control involves the internal relation between agent and action, a relation such that the action is an "outflowing of" the agent,[20] it is something that *she* does, bearing somehow her personal mark on it.[21]

---

question for which I have no answer. Maybe the concept of action entails some non-zero degree of attributability so that it makes no sense to refer to a ϕ-ing as an *action* of S's if the control condition is not met. Perhaps, as Davidson suggests (see his "Agency", n. 6), we should say that ϕ is something S *does*—a deed of hers—but not an *action* of hers. Again, I have nothing interesting to say in this respect.

[18] See John Martin Fischer & Mark Ravizza, *Responsibility and Control* (Cambridge: Cambridge University Press, 1998) chapter 2; Fischer, n.11, 101; Timothy O'Connor, "Indeterminism and Free Agency: Three Recent Views," *Philosophy and Phenomenological Research* 53 (1993), 499, 500.

[19] Fischer and Ravizza use the expressions "regulative control" and "guidance control", instead of alternative-possibilities control and agent control. See Fischer & Ravizza, n. 18, especially chapter 2.

[20] O'Connor, n. 18, 500.

[21] As John Martin Fischer nicely puts it, if acting responsibly means making a difference to the world, we may understand these two kinds of control with the following idea: "When an agent lacks alternative possibilities, he does not appear to make a *difference* to the world. And when he lacks agent control, *he* does not appear to make a difference to the world." Fischer, n.11, 101.

I am inclined to accept the semicompatibilist view that it is only the second kind of control (rather than both) that is necessary for responsibility for actions. But I need not argue for that here. My interest in elucidating the control condition is limited to what may impact on the judgments of relative blameworthiness in Success-Failure pairs. And, even if alternative-possibilities control is necessary, this will bear no consequence on the *relative* blameworthiness of the individuals in Success-Failure pairs—either both or none of them will lack alternative-possibilities control. The same, I shall show, is not the case with agent control. I focus hence on agent control.

The account of agent control I favor goes in terms of responsiveness to reasons: an agent S controls an action ϕ of hers if and only if S displays (some degree of) responsiveness to reasons in ϕ-ing.[22] On this account, very roughly, we are in control of something

[22] See Fischer & Ravizza n. 18; Joseph Raz, "When We are Ourselves: The Active and the Passive," in *Engaging Reason* (Oxford: Oxford University Press, 1999), 5-21. Raz, however, claims that the criterion of reasons-responsiveness, which he offers as an account of the distinction between "the active" and "the passive" in human life, is not a condition of responsibility. Indeed, he states that we may be responsible "for actions which we rightly disavow as not being 'us'. When such actions are wrong, responsibility for them may be attenuated and may not amount to guilt. But it may well give rise to duties to make amends, the sort of duties which only those responsible for an action or for a failure to act have." (p. 6) The apparent disagreement between my argument and this point of Raz's hides a mere semantic difference. As I'm using the term here, responsibility stands for susceptibility to moral responses (response-worthiness). When this response-worthiness is for wrongdoing, being responsible amounts to being guilty. Of course there are broader uses of the concept of responsibility, and Raz is giving to the term one such use. For example, we could understand "responsibility" in terms of what T. M. Scanlon called "accountability," so that one is responsible for, say, an action if one may be properly called on to give an account of why one performed that

that we do—where the "something" is our action being such that it satisfies a given description—when it is our response to the circumstances as we see them. What we thus do is *our* doing in the strong sense of attributability because, and to the extent that, it expresses what things we take to be reasons for action, how we weigh them, and how we translate them into action—it expresses, in other words, what may be called our practical identity.

Reasons-responsiveness is a modal concept—its meaning is not fixed by what is the case in the actual world. The attributability of S's ϕ-ing in the actual world as S's own depends on S's behavior in a suitable range of possible worlds or scenarios. In John Martin Fischer's and Mark Ravizza's rendition of the account, S's ϕ-ing is reasons-responsive only if, holding fixed the kind of action-producing mechanism that issues in ϕ-ing in the actual world, there exists some possible world in which there is a sufficient reason to not-ϕ, and the agent does not-ϕ for that reason.[23] Two remarks are in order.

First, Fischer and Ravizza restrict the set of possible worlds within which to check for reasons-responsivenes to those where the kind of "action-producing mechanism" at play in the agent's doing ϕ in the actual world is hold fixed.[24] The concept of an "action-

---

action. The account the agent gives may be such that it makes a moral response inadequate. T. M. Scanlon, "The Significance of Choice," *The Tanner Lectures on Human Values* (Oxford, 1986), 171; also Fischer, n. 11, 95. Under this broader notion of responsibility as accountability, I think Raz's claim is correct.

[23] See Fischer & Ravizza, n. 18, chapter 2, and 63-4.

[24] Id., 38-9.

producing mechanism" refers to the process that leads to the action, or "the way the action comes about,"[25] and was introduced to help deal with Frankfurt style counterexamples to the requirement of alternative-possibilities control. Harry Frankfurt's well known, and much discussed, counterexamples[26] involve an agent, Jones, who decides to ϕ after careful deliberation, and consequently does ϕ intentionally. The case also involves a second agent, Black or the counterfactual intervener, who fervently wants that Jones ϕ-s, is able to predict whether Jones is about to decide *not* to ϕ, and also to manipulate Jones's brain and nervous system to make her decide to ϕ, and to do ϕ accordingly—we are asked to imagine that Black has secretly placed a device in Jones's brain that allows him to remotely monitor and operate Jones's brain activity. If Jones were about to decide not to ϕ, Black would intervene and just make her decide to ϕ. Thus, there are two possible scenarios. In the first, and actual, scenario Jones decides to ϕ by herself, and does actually ϕ. In the second scenario, the counterfactual scenario, Jones also does ϕ but as a result of Black's intervention. These two scenarios test our intuitive grasp of the concept of action producing mechanism. For both scenarios diverge from each other in the action producing mechanism issuing in Jones's ϕ-ing: in the actual scenario, the mechanism is Jones's normal practical reasoning, which involved some mental events, like some sort of intention to ϕ, some beliefs as to how to ϕ, etc., and some physical events associated with

[25] Id., 38. See also 46-7 for some precisions.

[26] See Harry G. Frankfurt, "Alternate Possibilities and Moral Responsibility," *Journal of Philosophy* 66 (1969), 829-39, at 835-6.

those mental events. In the counterfactual scenario, in contrast, the mechanism at play is different in kind, for it involves the external manipulation of the brain through Black's device.[27] The mechanism concept is admittedly vague, but it is precise enough for my argument.

(Besides holding fixed the act-producing mechanism, we should also restrict the inquiry to those possible worlds that have the same natural laws as the actual worlds has.[28])

---

[27] Fischer's and Ravizza's mechanism-based approach contrasts with a so-called agent-based approach in the following way. The intuitive reaction to Frankfurt style examples is that Jones is responsible for ϕ-ing, even though she cannot do other than ϕ-ing—that is why they are counterexamples to the requirement of alternative-possibilities control. Now if Jones is responsible for ϕ-ing, then she must be sensitive to reasons in ϕ-ing. Here is where Fischer's and Ravizza's mechanism-based approach comes in. Under the opposite, agent-based approach we would check for reasons-responsiveness without holding fixed the mechanism that issues in Jones's ϕ-ing in the actual world. Every relevant possible world in which there is sufficient reason for Jones not to ϕ will thus be a world in which Black would assure that Jones ϕ-s—either because she decides to ϕ by herself, or because Black intervenes when she is about to decide not to ϕ—and hence we will find no reasons-responsiveness, and therefore no responsibility (i.e., the counterintuitive answer). When we shift to the mechanism-based approach we remove from the relevant set of possible worlds every world in which the mechanism operating in Jones's ϕ-ing in the actual world is not at work—among them, every world in which Black preempts its operation through his manipulation of Jones's brain. Thus, if Jones is a normal person whose ϕ-ing in the actual world is produced by the normal exercise of her practical reasoning, there will be some world in which no other mechanism is at work, there is sufficient reason for Jones not to ϕ, and Jones does not ϕ. Hence the mechanism-based approach accommodates the intuitive reaction to Frankfurt style cases.

[28] Fischer & Ravizza, n. 18, 44.

Second, the kind of reasons-responsiveness just introduced is too weak. There are cases of indisputably non-responsible agents which exhibit this weak sensitivity to reasons. The behavior of morally incapable individuals may very well exhibit this sensitivity to reasons. Fischer and Ravizza propose the example of a crazy assassin who kills all the passengers on a ferryboat, and who does the same in all of the relevant possible worlds except in one in which a passenger is smoking a Gambier pipe. In that possible world, the assassin takes the smoking of the pipe to be a reason for him not to slay the people on board, and then refrains from killing the passengers for that reason.[29] The assassin is of course insane and hence not responsible; and yet his action-producing mechanism is, though poorly, reasons-responsive. What responsibility requires, Fischer and Ravizza then add, is a more demanding kind of responsiveness, one that is revealed in the *range* of relevant possible worlds in which the agent recognizes there is a reason not to ϕ, and does not ϕ for that reason. To warrant responsibility, that range of possible worlds should evince an intelligible pattern of reasons that the agent is receptive to, a pattern that includes moral reasons.[30]

The relevant criterion for reasons-responsiveness should then be adjusted as follows: S's ϕ-ing is reasons-responsive if and only if, holding fixed the operation of the kind of action-producing mechanism issuing in ϕ-ing in the actual world, (i) there exists a range of possible worlds in which there is sufficient reason to not-ϕ, and the agent does not-ϕ

[29] Id., 65-6.

[30] Id., 69-85; for discussion, see Gary Watson, "Reasons and Responsibility," *Ethics* 111 (2001), 374-94, reprinted in Gary Watson, *Agency and Answerability* (Oxford: Clarendon Press, 2004), 289-317.

for that reason, and (ii) that range of possible worlds evinces an intelligible pattern of reasons (some of which are moral reasons) that the agent is receptive to.

Fischer's and Ravizza's criterion of reasons-responsiveness captures only one aspect or dimension of ownership. This dimension of ownership is, in a sense, negative, as it is represented by the agent doing *other* than ϕ-ing in the relevant set of possible worlds. I call it, following Fischer and Ravizza, *reactivity*.[31] There is also, I contend, a positive dimension of reasons-responsiveness that Fischer and Ravizza fail to explore, and which is crucial to my argument for the orthodox view. This dimension reflects the agent's ability to ϕ out of reasons for ϕ-ing.

I use the term *reliability*, which I draw from Alfred Mele's and Paul Moser's work on intentional action, to refer to this positive dimension of agent control. Mele and Moser believe that a minimum of reliability is required for something that an agent does to be an *intentional* action of hers. The following is an example of an action that fails along the reliability dimension while meeting all the other conditions of intentional action:

[31] Fischer & Ravizza, n. 18, 41-2, 69-76. Fischer and Ravizza distinguish reactivity (as the capacity to not-ϕ out of reasons to not-ϕ) from *receptivity* (understood as the capacity to recognize reasons in the world). They argue that responsibility requires what they call *moderate* reasons-responsiveness, or MRR. MRR is the case for S's ϕ-ing if and only if there is *some* possible world (within the relevant set) in which there is reason for S to not-ϕ and S does not-ϕ for that reason (or weak reactivity), *and* there is a suitable *range* of possible worlds in which S *recognizes* that there are sufficient reasons not to ϕ (or strong receptivity). I need not make that distinction here.

> A nuclear reactor is in danger of exploding. Fred knows that its exploding can be prevented only by shutting it down, and that it can be shut down only by punching a certain ten-digit code into a certain computer. Fred is alone in the control room. Although he knows which computer to use, he has no idea what the code is. Fred needs to think fast. He decides that it would be better to type in ten digits than to do nothing. Vividly aware that the odds against typing in the correct code are astronomical, Fred decides to give it a try. He punches in the first ten digits that come to his head, in that order, believing of his so doing that he "might thereby" shut down the reactor and prevent the explosion. What luck! He punched in the correct code, thereby preventing a nuclear explosion.[32]

In this case, the shutting down of the reactor that obtains is something Fred does, something moreover that he does guided by the intention to shut down the reactor, and still it is not an intentional action of his—it is, as Mele and Moser put it, "too coincidental to count as intentional."[33] Unlike Mele and Moser, I'm not interested in analyzing the concept *intentional action*, but in elucidating a dimension of attributability. I take the case of Fred and the nuclear reactor to be a good example of a doing that is a poor expression of Fred's agency, that is, of a very low degree of attributability. I contend that we are inclined to believe that the shutting down of the reactor is not an *intentional action* of Fred's because we are inclined to believe that it is not, under that description, an *action*

[32] Alfred R. Mele & Paul K. Moser, "Intentional Action," reprinted in *The Philosophy of Action* (Alfred R. Mele ed., Oxford: Oxford University Press, 1997), 224.

[33] Id., 225

of Fred's—that is, that it is too coincidental to be his own doing in the strong sense of attributability.

Whatever the minimum of reliability might be for a doing to be an expression of agency, differences along the dimension of reliability determine differences in attributability. Suppose a companion case that is as close to Fred's case as it can be but for the fact that the agent, Fred*, does know what the ten-digit code is—for, say, he is the person in charge of the reactor and has access to the secret document containing the code. In the companion case, Fred*'s shutting down of the reactor is certainly his own doing. It is also meaningful to say, I contend, that Fred*'s shutting down of the reactor is *more* Fred*'s doing than Fred's shutting down of the reactor is Fred's own. There is, as it were, more of Fred*'s in the shutting down of the reactor in the companion case than there is of Fred's in the shutting down of the reactor in the original example. This difference in ownership may be accounted for as a difference in the relative degrees of reliability of the corresponding action-producing mechanisms.

Reliability is a variable property which may be understood, like the reactivity dimension of reasons-responsiveness, in terms of the agent's performance in a set of possible worlds. Something like the following criterion will do:

> S's ϕ-ing exhibits reliability if and only if, holding fixed the operation of the kind of action-producing mechanism issuing in ϕ-ing in the actual world, (i) there exists a number of possible worlds in which there is sufficient reason to ϕ, and the agent does ϕ for that reason. (ii) Degree of reliability is determined by the ratio of the

number of worlds in subset (i) to the number of worlds in the reference set of possible worlds in which there exists sufficient reason to ϕ.

This criterion captures the intuitive view that whereas Fred*'s shutting down the reactor is Fred*'s own doing in the strongest sense, Fred's shutting down the reactor is hardly an expression of his agency. What accounts for this difference in attributability is that the mechanism issuing in Fred's shutting down the reactor in the actual world (i.e., typing in the first ten digits that come to mind) is such that Fred shuts down the reactor in too narrow a range of possible worlds—in those in which the ten-digit code happens to be the same as the first ten numbers that come to Fred's mind. Whereas Fred*'s mechanism (i.e., typing in the ten digits that, he knows, figure as the code in the pertinent secret document) is such that Fred* shuts down the reactor in most possible worlds within the relevant set.

Since all other things are equal between these two cases, such difference in the reliability dimension warrants differential responsibility judgments—I venture to say that while Fred* may be praiseworthy for shutting down the reactor (and preventing the explosion), Fred may deserve praise only for trying to shut down the reactor: the attributability of his shutting down the reactor as his own is too close to zero. [34]

[34] As the contrast between these two cases show, reliability (and hence agent control) may depend on knowledge (or justified belief). But it may also depend on skill. Take my archery examples. It may be very well the case that the beliefs involved in my hitting the bull's eye in my first shot—in aiming at the target and releasing the arrow, say—are the same as the instructor's beliefs involved in her hitting the bull's

Success-Failure pairs: assessing relative ownership

Let's go back to Success-Failure pairs. Alvarez is blameworthy for α-ing, and her blameworthiness is determined by the amount of wrong done by her α-ing and the degree to which the α-ing may be attributed to Alvarez as her doing. The same goes for Borges, whose blameworthiness for her β-ing is a function of the amount of wrong done by β-ing and the attributability of this action as Borges's own doing. If, as I have assumed, α-ing is somehow more wrongful than β-ing is, then for the orthodox view to be true there must be a proportional difference in the attributability variables such that it outweighs the difference in the amount of wrong done by one and the other agent.

In this section I argue that there is indeed a difference in the relevant attributability variables between the cases in Success-Failure pairs that is similar in kind to the difference in attributability between cases like Fred's shutting down the reactor and Fred*'s shutting down the reactor, and between the archery instructor's hitting the bull's eye and my hitting the bull's eye. More specifically, I will maintain that we may conceive of the degree of attributability of Avarez's α-ing as her doing as a determinate fraction of the measure of the degree of attributability of Borges's β-ing as Borges's doing.

---

eye. And still there is an obvious difference in terms of reliability, and hence in agent control. That is what we call skill. The instructor is a skilled archer, which means that when acting from her normal action-producing mechanism, she exhibits high reliability measures in making her actions meet the archery relevant reasons for action—like the reasons there might be for hitting the bull's eye. See id., 252-3. For Mele & Moser's notion of skill, see id., 246.

The degree of attributability of an action, I claimed above, is determined by the degree of reasons-responsiveness exhibited by the agent in her action, which in turn has two aspects or dimensions—reactivity and reliability. There is no difference of attributability between the *Success* case and the *Failure* case as regards the dimension of reactivity. We measure reactivity, let's recall, by checking for possible worlds where the agent has sufficient reason to act differently than she does in the actual world, and does act differently for that reason. Reactivity does not vary from *Success* to *Failure* for the set of possible worlds in which Alvarez has a reason to not-α and does not-α for that reason, is identical to the set of possible worlds in which Borges has a reason to not-β and does not-β for that reason. This point is secured by the following two features of Success-Failure pairs.

The first feature is the fact that Borges and Alvarez are only numerically distinct—they are in every other respect one and the same person, and so they are identically prone to find motivation in the same sort of facts. More specifically, for any fact *f* and any act ϕ, if *f* is a reason for Alvarez to ϕ, it is also a reason for Borges to ϕ, and vice versa. Therefore, in every world in which Alvarez finds a reason to ϕ, and does ϕ for that reason, Borges would identically find a reason to ϕ, and do ϕ for that reason—and vice versa.

The second feature is the relation between actions α and β. Since β is the actional means to α (i.e., it is more basic than, and practically sufficient to, α), they are identical at the level of reasons. This identity at the level of reasons for α and β stands for the conjunction of the following four propositions: For any fact *f* and any agent S, (i) if *f* is a reason for S to α, it is also a reason for S to β; (ii) if *f* is a reason for S to β, it is also a

reason for S to α; (iii) if *f* is a reason for S to not-α, it is also a reason for S to not-β; and (iv) if *f* is a reason for S to not-β, it is also a reason for S to not-α. Propositions (i) and (iv) follow from the fact that β is a constitutive part of α, which is part of the meaning of its being more basic than α; and (ii) and (iii) follow from the relation of practical sufficiency, according to which one can't do β without making it likely that the β-ing turns out an instance of α-ing.

One may dispute the identity at the level of reasons between an action an its actional means, by thinking of the relationship as equivalent to the relationship between succeeding and trying to succeed. John Gardner has argued that there are reasons to succeed that are not as well reasons to try, as in the case in which I have reason to save a person from drowning, but there are no means available for me to try to save her.[35] If Gardner is right (which I doubt), the set of worlds in which we have reason to succeed is not identical to the corresponding set of worlds in which we have reason to try.

The relationship between an action and its actional means is not, however, the same as the relationship between succeeding and trying, in the sense Gardner uses these terms. In this sense there is something like succeeding without trying, whereas there is no α-ing without β-ing. Suppose again that "I α" stands for "I intentionally kill my colleague next door by throwing a grenade into her office." My actional means to that action is, say, my throwing a grenade into my colleague's office with the intention to kill her—so, in this example, "I β" stands for "I throw a grenade into may colleague's office with the inten-

[35] See John Gardner, "The Wrongdoing that Gets Results", *Philosophical Perspectives* 18 (2004), 53-88.

tion to kill her." The identity at the level of reasons that I'm claiming holds between actions α and β under those descriptions. Thus, the claim entails that if there is reason for me (not) to intentionally kill my colleague by throwing a grenade into her office, there is reason for me (not) to throw a grenade into my colleague's office with the intention to kill her. (And conversely, if there is reason for me [not] to throw a grenade into my colleague's office with the intention to kill her, there is reason [not] to intentionally kill my colleague by throwing a grenade into her office.)[36]

The two features of Success-Failure pairs I've just highlighted—Alvarez-Borges personal identity, and identity at the level of reasons between actions α and β—secure then that the set of worlds in which Alvarez finds a reason to not-α, and does not-α for that reason is identical to the set of worlds in which Borges finds a reason to not-β, and does not-β for that reason. Hence my point that there is no difference to draw between Success and Failure along the reactivity dimension of attributability.

I turn now to the reliability dimension of attributability. Unlike reactivity, reliability does mark a difference in attributability of Alvarez's α-ing as her doing, and Borges's β-ing as her doing. I analyze reliability of the agent's doing a given action, let's recall, as the frequency of success in doing that action across a range of possible worlds. The

[36] This is not to deny that my α-ing in this example might be re-described so that it does not encompass my β-ing, and in a way that makes sense of the claim of the disidentity between reasons to succeed and reasons to try. For instance, it is true in my example that I kill my colleague, and the reasons I have not to kill her exceed the reasons I have not to β or, more generally, not to *try* to kill her—for I do what those reasons would have me not to do even if I kill my colleague without trying to kill her (say, by accident).

measure of reliability is thus to be expressed as a value ranging from one, for full reliability, to zero, for total unreliability, depending on the frequency with which the pertinent action obtains within the relevant set of possible worlds.

Let r_β express the degree of attributability (as a function of reliability) of Borges's β-ing in *Failure*. It will be at some point between full attributability (where $r_\beta = 1$) and non-attributability (where $r_\beta = 0$), depending on (i) the number of relevant possible worlds in which there is sufficient reason for Borges to β, and Borges does β for that reason; and (ii) the number of relevant possible worlds in which there is sufficient reason for Borges to β. The value of r_β represents the ratio of the number in (i) to the number in (ii).

Whatever the actual value of r_β is, the value of the corresponding variable expressing the degree of attributability (as a function of reliability) of Alvarez's α-ing in *Success* (or r_α) is equivalent to a fraction of r_β. The following three features of Success-Failure pairs support this claim.

First, r_α stands for the ratio of (a) the number of relevant possible worlds in which there is sufficient reason for Alvarez to α, and Alvarez does α for that reason, to (b) the number of relevant possible worlds in which there is sufficient reason for Alvarez to α. Now, it follows from the fact that β is the actional means to α, that Alvarez does α only in those worlds in which she does β. Moreover, since it is part of the definition of Success-Failure pairs that there is some room for failure in Alvarez's α-ing—for it is within that space where the *Failure* case lies—then it follows that the set of the worlds in which Alvarez does α is a subset of the set of worlds in which she does β. There will be some

possible world (or worlds), that is to say, in which Alvarez does β without as well doing α. So, the measure of variable r_α will be lower than the corresponding measure of the frequency of Alvarez's β-ing within the reference set of possible worlds relevant to reliability for Alvarez's α-ing—that is, the set of relevant possible worlds in which there is sufficient reason for Alvarez to α.

Second, the identity at the level of reasons of α and β allows us to move from that observation to the claim that the measure of the attributability of Alvarez's *β-ing* as her doing is higher than the measure of the attributability of Alvarez's *α-ing* as her doing. Indeed, we assess the reliability variable for Alvarez's β-ing in *Success* by estimating the frequency of Alvarez's β-ing within the set of relevant possible worlds in which there is sufficient reason for Alvarez *to β*. The identity at the level of reasons of α and β warrants that this reference set is identical to the reference set within which to measure the variable r_α (i.e., the set of possible worlds in which there is sufficient reason for Alvarez *to α*). Since the frequency of Alvarez's β-ing within that set is higher than the frequency of her α-ing within that set (for there are some worlds in which she does β without thereby doing α), it follows that the value of the variable expressing the attributability of Alvarez's β-ing in *Success* is higher than r_α.

Third, cases in Success-Failure pairs are as close to each other as to secure that the β-ing in *Success* is as Alvarez's doing as the β-ing in *Failure* is Borges's doing. Indeed, we may conceive of *Success* and *Failure* as describing two members of the reference set of possible worlds within which we measure the frequency of Alvarez's α-ing (and β-ing) to

establish attributability ("Alvarez" and "Borges" are the names with which we refer to one and the same agent, depending on the world she is in, a Success world or a Failure world). Hence, the degree of attributability of the β-ing in *Success* to Alvarez as her doing is equivalent to r_{β}, which is in turn, then, higher than r_{α}.

Success-Failure pairs illustrate a general thought about attributability. The thought is that attributability decreases as the complexity of the action increases, all other things being equal. More specifically, for any agent S, and any non-basic action ϕ of hers, there will be some more basic (or less complex) action that S does as a part of ϕ by which S does ϕ—e.g., S turns the light on by flipping the switch; her flipping the switch is more basic than her turning the light on. The more basic the action is, the fewer the external conditions that need be met for it to succeed—e.g., S succeeds in flipping the switch even if the bulb fails to light, but she needs the bulb to actually work for it to be the case that she turns the light on. Hence, if there is room for failure in the more complex action—e.g., if it is possible that the bulb fails—the set of the possible worlds in which the agent does the more complex action will be smaller than the set of the corresponding possible worlds in which she achieves the more basic action. (In the extreme, attributability would reach its maximum level with the corresponding, non-comparatively basic act.) The difference in the attributability degrees in Success-Failure pairs instantiates this general thought—that is, r_{β} is higher than r_{α} because β is more basic than α, and there is no difference between Alvarez's β-ing in *Success* and Borges's β-ing in *Failure*.

This argument shows that the level of attributability of Alvarez's α-ing is necessarily lower than the corresponding degree of attributability of Borges's β-ing. Let me now state more clearly how to measure such difference.

Let r_α be the coefficient representing the degree of Alvarez's responsibility for her α-ing., and r_β the corresponding coefficient representing the degree of Borges's responsibility for her β-ing, and reduce both coefficients (r_α and r_β) to the corresponding reliability variables.

According to what I have argued, r_β may be understood as the measure of the frequency with which the agent does β within the relevant reference set of possible worlds in which there is sufficient reason to β. Coefficient r_α, in turn, represents the measure of the frequency with which the agent does α within the same reference set—i.e., due to the identity at the level of reasons between α and β, the set of the worlds in which there is sufficient reason to α is identical to the set of the worlds in which there is sufficient reason to β. Since the agent can only do α by doing β as a part of it, the set of the worlds in which the agent does α is in turn a subset of the set of the worlds in which she does β. Let variable $r_{\alpha|\beta}$ represent the ratio of the number of α's to the number of β's within the reference set of possible worlds, or the fraction of β's which are also α's within that set. Variable $r_{\alpha|\beta}$ thus represents the degree to which the attributability of Alvarez's α-ing is lower than the corresponding attributability of Borges's β-ing. Indeed, if r_β is the frequency of the β's obtaining within a given set of possible worlds, r_α is the frequency of the α's obtaining within the same set, and the frequency of α's depends on the frequency

of β's (for an instance of α can only obtain if an instance of β has obtained, but not the converse), then applying elementary probability calculus we may calculate r_α as the product of r_β times the relative frequency of α's to β's, or $r_{\alpha|\beta}$. That is, $r_\alpha = r_\beta \times r_{\alpha|\beta}$. Variable $r_{\alpha|\beta}$ is therefore a sort of discounting coefficient which multiplied by r_β yields the value of r_α. In that sense I claim that $r_{\alpha|\beta}$ represents the degree to which the attributability of Alvarez's α-ing is lower than the corresponding attributability of Borges's β-ing.

To complete my argument for the orthodox view, then, I shall offer next a conception of wrongness such that the value of the wrong done by Borges's β-ing is lower than the wrong done by Alvarez's α-ing in a degree that matches the difference in attributability as represented by variable $r_{\alpha|\beta}$.

Discounted wrongness

I contend that the most natural account of the wrongness of the actional means to a wrongful action—that is the wrongness of an action that stands to a wrongful action as β stands to α, involves the following two claims. First, the wrongness of the actional means is of a derivative nature: the actional means is wrongful, that is, because the less basic action to which it is a means is wrongful. Second, the path through which this derivative wrongness is derived is some sort of likelihood or probability of the actional means becoming the less basic, non-derivatively wrongful action. For the actional means β, and the wrongful action α to which β is the actional means, this natural view may be put in the following terms. If W_α represents the amount of wrong done by α-ing, then the

wrongness of β-ing, or W_β, is a function of W_α and the conditional probability of the agent achieving α once she has done β. Let *p* represent the measure of this conditional probability (ranging between 0, where β-ing creates no probability of α-ing, to 1, where β-ing necessarily ends in α-ing), and therefore $W_\beta = p \times W_\alpha$. Under this view, then, the wrongness of an actional means β to a wrongful action α varies with the probability of the agent doing α once she has done β. The likelier it is that the β-ing turns out an α-ing, the more wrongful it is for the agent to do β. In the extremes, where doing β has no chance whatsoever to bring about the wrongful action α, it is not derivatively wrongful for the agent to β; and where doing β brings about α necessarily, it is no less wrongful for the agent to β than it is for her to α.

Probability in this account of derivative wrongness is neither objective, physical probability, nor purely subjective probability understood as degree of confidence. Indeed, we may speak in terms of probability, as this term figures in the account, even if the world is deterministic. We say, for instance, that I create a probability *p* (where *p* is at some point higher than 0 and lower than 1) of killing you by playing Russian roulette on you, or that shooting at you after checking that the gun is fully charged creates a higher probability of killing you than just playing Russian roulette on you, whatever the actual outcome happens to be, and even if it is fixed by the physical history of the world and its natural laws. To claim in this sense that a particular event makes it probable that another particular event ensues is to express a kind of uncertainty about how the world is, or how it functions. This is not to say, on the other hand, that the language of probability, as it is used in this account of wrongness, is a purely subjective matter. In playing Russian roulette on

you I act derivatively wrongly because by playing Russian roulette on you I may very likely kill you, and killing you is non-derivatively wrong. Now, in playing Russian roulette on you I wrong you in this sense even if, being unreasonably confident in my good luck, I in fact believe that you will not die as a consequence of the shot. By the same token, I do no wrong in trying to kill you by sticking pins in a voodoo doll because there is no chance whatsoever that I kill you by thus sticking pins in such a doll, even though I unreasonably believe that by sticking pins in this doll I make it likely that you die.

Particularly, we may understand the probability variable *p* as a measure of the frequency with which the non-derivatively wrongful action α obtains within a reference set of possible worlds in which the agent does β as she does in the actual world. The reference set should be so restricted as to reflect the relevant kind of uncertainty that judgments of probability are meant to express. Suppose that S punches you in the nose and your nose starts bleeding, and we want to estimate the probability of S's making your nose bleed by punching you in the nose. If we consider what is the frequency of S's making your nose bleed by punching you in the nose within the set of worlds which are identical to the actual world in every causally relevant respect up to the time where S punches you in the nose, then we will come up with a frequency that matches the objective, physical probability or, if the actual world is deterministic, with the finding that S makes your nose bleed by punching you in the nose in every possible world within the set. So, in order to capture the kind of probability in force in discounted wrongness we should adjust the restriction of possible worlds, moving from such causally relevant identity to the actual world to something like perceived identity from a given agential point of

view. Call this latter set of possible worlds, the set of possible worlds *practically equivalent* to the actual world. Practical equivalence is relative to a given kind of action, and an agential standpoint. Thus, possible world *w* is practically equivalent to the actual world as regards the action "making your nose bleed by punching you in the nose" and from S's point of view if and only if S finds at *w* as propitious an opportunity to make your nose bleed by punching you in the nose as she finds in the actual world. Put another way, *w* is practically equivalent to the actual world with respect to that kind of action, and S's point of view, because if S wanted to make your nose bleed by punching you in the nose, she would be totally indifferent whether she is at *w* or in the actual world.

Practically equivalent worlds may be very different from one another. They may differ both in features epistemically unavailable to the reference agent—like low level physical properties, or the contents of secret documents—and in any accessible feature bearing no influence on the prospects of the action from the point of view of the reference agent.

By restricting the reference set of possible worlds to the worlds being practically equivalent to the actual world we make the probability judgment reflect the epistemic limitations of the reference agent (i.e., the agent whose epistemic abilities we take into account to determine practical equivalence), so that the more epistemically able the reference agent is, the closer to the physical, objective probability (if there is such a thing) the resulting probability measure will be.

It is unclear to me who the relevant reference agent should be, whether the agent whose action is under evaluation or someone else—perhaps something like an ideal, "rea-

sonable" agent—though I am inclined in favor of the former. The choice of the actual agent's stance as the relevant agential point of view is in keeping with the theory that moral reasons in general, and wrong making factors in particular, must fall within the cognitive capacity of the agent whose conduct they purport to guide. In general, judgments of rightness or wrongness, Jonathan Dancy has argued, are relative to a certain body of facts; and judgments of moral rightness or wrongness, in particular, are relative to the body of facts available to the agent. The reason is that if something is to count to us as a moral reason, it must be "capable of being practically relevant for us."[37] By including the agent's point of view in the criterion of practical equivalence of possible worlds, we secure that the resulting wrong making factor—i.e., the probability of the action turning out a making of your nose bleed—is accessible to the agent and, therefore, may play a role in guiding her conduct.

Before defending the basics of this view of the wrongness of the actional means to a wrongful action against its possible alternatives, I wish to complete my case for the orthodox view by arguing that in Success-Failure pairs the degree to which the attributability of Alvarez's α-ing is lower than the corresponding attributability measure of Borges's β-ing is equivalent to the degree to which β-ing is less wrongful than α-ing is—or why, formally, $p = r_{\alpha|\beta}$.

Both $r_{\alpha|\beta}$ and p may be understood as expressing the ratio of the number of α's to the number of β's within a suitably restricted set of possible worlds. Variable p expresses the

[37] See Jonathan Dancy, *Practical Reality* (Oxford: Oxford University Press, 2000), 56-60 (quoted phrase at 59).

ratio of the number of times the agent does α to the number of possible worlds in the set of possible worlds that are practically equivalent to the actual world with respect to α and from the agent's point of view (the actual world being that in which *Failure* obtains) and in which the agent does β. Call this reference set, *P*. The $r_{\alpha|\beta}$ variable, on the other hand, stands for the ratio of the number of times the same agent does α for a reason to α to the number of possible worlds in the set of possible worlds in which there is sufficient reason for the agent to α, and the agent does β for that reason. Call this latter set, *R*.

These two sets, *P* and *R*, are identical to each other—and hence $r_{\alpha|\beta}$ and *p* state identical ratios—if the features of the agent's doing that the α label picks out are such that (i) the agent can only do α for a reason, and (ii) doing α entails that the agent (in doing β) entertains causal beliefs on the prospects of her achieving α of some given kind—so that if acted with relevantly different beliefs, if would not be α (nor β) that she does, but something else. Indeed, condition (i) secures that in every world in which the agent does α (or just its actional means, β) she does so in response to a reason to α, and so that every world in *P* belongs as well in *R*. Condition (ii), in turn, warrants the converse, that is, that every world in *R* also belongs in *P*; for every possible world in which the agent does α (or just β) for a reason is a world that is practically equivalent to the actual world with respect to α and from the agent's point of view.

Conditions (i) and (ii) are met, I contend, in every case of intentional wrongdoing that may play the role of a *Success* case. This turns on two assumptions. The first assumption

is that when an agent does something intentionally, she does it for a reason.[38] So, if "α" describes an intentional action (i.e., an action that is intentional under the α description), there is no world in which the agent does α (or its actional means β) without doing it for a reason. Hence, condition (i) is met.

The second assumption is that, if "α" describes an intentional action, it implies that the agent acts (in doing β) with certain beliefs as to the prospects of her achieving α.[39] So, every possible world in which the agent does α (or just β) is a world in which she entertains (in doing β) those same beliefs as to the prospects of her achieving α, and also a world in which the *grounds* of those beliefs are relevantly the same. By "grounds" I mean the features of the world on the basis of which the agent believes that she will (or may with a given degree of likelihood) achieve α once she has done β. This secures that every feature of the actual world which from the agent's point of view is causally relevant for her achieving α must obtain in every possible world in which the agent does α (or just β). In other words, every possible world in which the agent does α (or just β) is practically equivalent to the actual world (in which the agent has done α) with respect to α and from the agent's point of view. Hence condition (ii) is satisfied.

---

[38] See for this relation between acting intentionally and acting for a reason, e.g., Donald Davidson, "Actions, Reasons, and Causes," reprinted in *Essays on Actiuons & Events* (Oxford: Clarendon Press, 1980), 6; Robert Audi, "Acting for Reasons," reprinted in *The Philosophy of Action* (Alfred R. Mele ed., Oxford: Oxford University Press, 1997), 75-105.

[39] See, e.g., Mele & Moser, n. 32, 240-1.

If my contention is right, variables $r_{\alpha|\beta}$ and p in Success-Failure pairs necessarily represent equivalent ratios, at least when that for which the agents are to blame—that is, the actions that "α" and "β" describe—are intentional actions. In the domain of those kinds of actions, then, the orthodox view that successful offenders are no more to blame than their failing counterparts will be true on the basis of the argument I have deployed.

In defense of discounted wrongness

Before concluding I would like to briefly defend my account of the wrongness of actional means to independently wrongful actions. I think that the account is at least as good as its alternatives. Consider first the view that denies that there is any wrong done by the actional means to a wrongful act: the wrongness of an action α, wherever it comes from, is not conveyed to its actional means. Take a version of the Success-Failure pair in which Alvarez and Borges attempt to kill Victor by shooting him; Alvarez succeeds, and Borges fails. Together with my assumption that blameworthiness entails wrongness, this view implies that agents like Borges, who attempt to do serious wrong but fail, are blameless. I take this conclusion to be highly implausible, and therefore a reason not to abandon my account in favor of this alternative view.

One may, however, escape the implausible conclusion by revising the initial assumption on the relation between blameworthiness and wrongdoing, thereby making room for judgments of blameworthiness in the absence of wrongdoing. Though possible, this move is not without costs. For it breaks the identity between that which we would say an agent is to blame *for*, and that *in virtue of which* she is to blame. To illustrate, take again

the version of the Success-Failure pair in which Alvarez and Borges attempt to kill Victor, Alvarez succeeds, and Borges fails. The cases describe an episode in the agents' life—their respective attempts to kill Victor—which leads us to the judgment that both agents are to blame. In judging an agent blameworthy we are expected to state what is that which the agent is to blame *for*. In my cases the natural answer, I take it, will be that Alvarez is to blame *for killing Victor intentionally*, and that Borges is to blame *for attempting to kill Victor*. The account that best accommodates what we do in filling in these "for" clauses is that we are stating the *ground*, or *basis* of the judgment of blameworthiness we have just passed; that is to say, we indicate the features of the case *in virtue of which* the agent is blameworthy. Under this account, what we offer as the basis of the judgment, in the "for" clause, should thus *explain* why the agent deserves blame. We honor this expectation if in "S is to blame for ϕ-ing" we imply that it is morally wrong for S to ϕ. For the fact that S's ϕ-ing is *morally wrong* would indeed account for the fact that it is *blame* that S deserves. (The fact that the ϕ-ing is *S's own doing* would account for the fact that she *deserves* blame.) Were S's ϕ-ing not wrong, if it were (say) a morally stupendous action, it would be misleading, if intelligible at all, to say that S is to blame for ϕ-ing.

My point, in sum, is that it makes sense to give our natural answer "Borges is to blame for attempting to kill Victor" precisely because it is wrong for Borges to attempt to kill Victor.[40] By denying that actions such as Borges's attempt to kill Victor can be wrong,

[40] This view of what it is for which a person deserves blame is defended by Copp, n. 12, 448-51.

we render the natural answer "Borges is to blame for attempting to kill Victor" meaningless.

Michael Zimmerman—who argues that blameworthiness does not depend on wrongness—posits that in stating that which the agent is to blame for we don't express the basis of the judgment of blameworthiness, but its *scope*.[41] Zimmerman also believes that attempts to do what is wrong are not themselves wrongful,[42] and therefore that actions such as Borges's attempting to kill Victor are permissible actions. Permissible actions cannot ground judgments of blameworthiness. So, if Borges is to blame, the feature of the case in virtue of which she is to blame must be something else. Zimmerman proposes that it is Borges's being such that she *would* do wrong (i.e., by killing Victor under certain circumstances) what makes her blameworthy.[43] But, if this is right, what are we doing when we say of Borges that she is to blame *for attempting to kill Victor*? Nonsense, seems to be Zimmerman's answer. Whether an agent is to blame, and how much blame she deserves are both determined by that which counts as the basis of blameworthiness, not by its scope. There is indeed what Zimmerman calls blameworthiness *tout court*, that is, blameworthiness with no scope whatsoever, or without anything at all *for* which the agent is to blame.[44]

---

[41] Zimmerman, n. 2, 560-1. He makes the point more generally in terms of "responsibility," but it also goes for blameworthiness as a species of responsibility.

[42] Michael J. Zimmerman, "A Plea for Accuses," *American Philosophical Quarterly* 34 (1997) 229-43.

[43] Zimmerman, n. 2, 565.

[44] Id., 563-5.

I conclude that if we hold to the intuitively plausible view that agents such as Borges in *Failure* are blameworthy, the natural way to account for this is to argue for the wrongness of what these agents do—i.e., attempting to do what would be wrong if done. The alternative consisting in freeing blameworthiness from the wrongdoing requirement fails to accommodate a feature of our moral discourse (the natural "for" clauses of our judgments of blameworthiness).

Another alternative to my account of the wrongness of actional means would be to concede that attempts to do what would be wrong if done are wrongful themselves, but to claim that what makes such attempts wrong is the agent's acting on an intention she ought not to act upon—rather than the probability that the non-derivatively wrongful act obtains. Alec Walen has argued that there are what he calls "illicit intentions," that is, intentions we should not form, or act upon, so that, when we act on an illicit intention we do wrong, even though the act itself that we produce would be permissible under another (licit) intention.[45] The wrongness of attempts, in Walen's view, may be accounted for in these terms. As he puts it, "[i]t is *because* an agent attempting to commit a crime is acting on an illicit intention that he is acting impermissibly."[46] An intention is illicit in this sense, according to Walen, if it makes the agent be "disposed to perform actions that can

[45] Alec Walen, "The Doctrine of Illicit Intentions," *Philosophy & Public Affairs*, 34 (2006), 39-67. Walen draws a distinction between wrongness and impermissibility that I need not make here. As I used the terms in this paper, "morally wrong" and "morally impermissible" are synonymous expressions.

[46] Id., 66-7.

independently be determined to be impermissible."[47] In a case like *Failure* (under the killing-Victor interpretation), Borges's intention is illicit because it is such that Borges would carry out such intention by killing Victor (in the absence of justifying circumstances), and killing Victor (in the absence of justifying circumstances) is wrong independently of what the intention of the killer is. Borges's attempt to kill Victor is wrong, then, in virtue of his acting on an illicit intention, and thus quite independently of what he does *to* Victor other than intending to kill him.

The illicit intention approach to the wrongness of attempts fails to capture what appears to be an important dimension of attempts, namely, the imposition of risk. Surely it makes a great deal of a difference whether Borges attempts to kill Victor by playing Russian roulette on him, or rather by sticking pins on a voodoo doll, even though in both cases Borges exhibits the same intention to kill Victor.[48] At any rate, it does make a lot of a difference in the criminal law,[49] and I see no reason for this legal difference not to be a reflection of a moral difference. To the very least, it is a counterintuitive claim that, all other things being equal, it is permissible for an agent to create on an non-consenting victim a one in six chance of an instantaneous death, as in the Russian roulette example. The plausible view, as Judith Thomson has convincingly argued, is that one ought not do

[47] Id., 39. This is a very rough statement of Walen's argument. See id., 50-6.

[48] In one and the other case, Borges has different beliefs, which leads him to engage in different courses of actions in order to carry out his intention to kill Victor. Given this intention to kill Victor and the different sets of believes, Borges is moved to form different more specific intentions regarding what to do in order to kill Victor. At the level of the chosen means, then, there will be different intentions.

[49] See, e.g., Dressler, n. 8, 370-2.

that, and so quite independently of one's actual beliefs and intentions.[50] I don't mean that this plausible answer is without problems, as Thomson so acutely remarks.[51] But the same goes for the opposite view that denies wrong-making nature to the imposition of risk.

My account of the wrongness of attempts like Borges's attempt to kill Victor is that it is wrong for Borges to attempt to kill Victor because she ought not to kill Victor, and attempting to kill Victor is Borges's actional means to killing him. Conceiving of the wrongness of actional means, and attempts in particular, along these lines is in keeping with the wrongness of risk-imposition. Indeed it is part of what makes of an action β the actional means to an action α that by β-ing the agent makes α-ing likely. When α entails that some evil outcome obtains, like harm to somebody else, we talk of that likelihood in terms of risk. Borges's attempting to kill Victor is wrong, then, because killing Victor is wrong, and by attempting to kill Victor Borges makes likely that he kills Victor—that is, because by attempting to kill Victor he imposes on Victor the risk of getting killed.

Conclusion

Let me summarize my argument. I started with the assumptions that in Success-Failure pairs the successful agent does more wrong than the unsuccessful counterpart does, and that the magnitude of wrong done impacts in the agent's degree of blameworthiness. I argued next that blameworthiness is not only a function of the magnitude of the wrong

[50] See Thomson, n. 5, 173-91.

[51] Id., 184-8.

done but also of the degree to which the wrongdoing is attributable to the agent as her own doing. Thus, applying the Nozickian framework r_ϕ x W_ϕ, we can understand the orthodox view in terms of the following equation,

(1) $$r_\alpha \text{ x } W_\alpha = r_\beta \text{ x } W_\beta.$$

The left-hand side of this equation expresses the magnitude of Alvarez's blameworthiness for his α-ing in *Success*, as it is determined by the value of the wrong done by α-ing (W_α) times the coefficient representing the corresponding degree of attributability (r_α), or r_α x W_α. On the right-hand side we find the expression of Borges's degree of blameworthiness for her β-ing in *Failure*, as it is correspondingly determined by r_β x W_β. The orthodox view is true, I claim, because the difference between attributability coefficients (r_α and r_β) outweighs the difference in wrong done by one and the other agent.

I interpret the attributability coefficients (r_α and r_β) in equation (1) to reflect only the measures of reliability of each agent's doing—the other dimension of attributability, i.e. reactivity, being constant from one case to the other. Thus, they are understood to express a measure, between 0 and 1, of the frequency with which the pertinent action obtains within a given set of possible worlds. Coefficient r_α in equation (1), I argued, is a fraction of coefficient r_β and may be reduced, I proposed, to the product r_β x $r_{\alpha|\beta}$, where $r_{\alpha|\beta}$ represents the ratio of the number of actions of the type α to the number of actions of the type β within the relevan set of possible worlds. Together with the claim that $r_\alpha = r_\beta$ x $r_{\alpha|\beta}$, equation (1) yields,

(2) $$r_\beta \text{ x } r_{\alpha|\beta} \text{ x } W_\alpha = r_\beta \text{ x } W_\beta,$$

which in turn yields the simpler equation,

(3) $$r_{\alpha|\beta} \text{ x } W_{\alpha} = W_{\beta}.$$

I maintained that the wrongness of Borges's β-ing should be conceived as a derivation of the wrongness of α-ing, and its measure as a fraction of the value of the wrong done by α-ing. More specifically, I argued that the value of wrong done by β-ing equals to the value of wrong done by α-ing discounted by the probability of the agent achieving α once she has done β. If we let p stand for a discounting coefficient representing the measure of that probability, then we may state the wrongness of Borges's β-ing as $W_{\beta} = p$ x W_{α}. Replacing W_{β} in equation (3) with its equivalent expression leads to,

(4) $$r_{\alpha|\beta} \text{ x } W_{\alpha} = p \text{ x } W_{\alpha},$$

which in turn yields,

(5) $$r_{\alpha|\beta} = p.$$

As (5) makes explicit, my argument ultimately depends on there being an argument that the frequency represented by coefficient $r_{\alpha|\beta}$ is equivalent to the measure p of the probability of the agent achieving α once she has done β. I argued that such is the case at least when the α label stands for an intentional wrongdoing.