Accidentally Doing the Right Thing

Zoë Johnson King

University of Michigan

Email: zoejk@umich.edu

Word count of paper: 12,613 words.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/phpr.12535

MS. ZOE ANNIS JOHNSON KING (Orcid ID : 0000-0001-5248-472X)

Article type : Original Article

Accidentally Doing the Right Thing

1. Introduction

This paper is about moral worth. Moral worth is a positive status that some, but not all morally right actions possess. There is a live dispute as to what makes the difference.

We can begin to get a handle on this dispute by considering two fictional characters. One is from a classic American novel by Mark Twain (1884). The other is from the movie *Star Wars: The Force Awakens*.

> FINN FROM STAR WARS: Stormtrooper FN-2187 was bred and trained to fight for the First Order, the current incarnation of the dark side of the force. But, unlike other stormtroopers, he has a conscience. On his first intergalactic mission he is shocked and appalled by the blood spilt and carnage wrought by his comrades. Later, he is assigned to guard Poe Dameron, a pilot for the resistance movement who has been captured by the First Order. Recalling the carnage of his intergalactic mission, FN-2187 chooses instead to rescue Poe and escape with him. On hearing of this plan, Poe asks, "Why are you helping me?", and FN-2187 – his face drenched in sweat and momentarily solemn – replies, "*Because it's the right thing to do*".

> HUCKLEBERRY FINN: Huckleberry ("Huck") Finn is a teenager growing up in the American South in the mid-1800s. Huck has absorbed the racist ideology of his contemporaries; he fully believes that slaves are the property of their owners, that helping a slave to escape is stealing, and that it is therefore morally wrong. Nonetheless, Huck befriends a fugitive slave named Jim. And when he gets the opportunity to report Jim to the authorities, he chooses instead to lie and thus help Jim to escape. Huck is profoundly conflicted at this point; he is convinced that what he is doing is morally wrong, yet he cannot resist the urge to help his friend.

In *Star Wars*, Poe later gives FN-2187 the nickname "Finn". So here we have two fictional characters, both named Finn. Their similarity extends beyond their names: each helps somebody who was unjustly held captive to escape, in a poignant moment of character development that is pivotal to their respective plots. And both agents thereby do something that is morally right. The question at issue in this paper is whether the two Finns perform actions with moral worth.

There are two main views in this dispute:

> **KANTIAN VIEW**: Someone performs an action with moral worth only if she is motivated to do the right thing by the very fact that it is right.

> **NEW VIEW**: Someone performs an action with moral worth if she is motivated to do the right thing by the features that make it right (the "right-making features").

The traditional Kantian view is that, for an act to have moral worth, the agent must do it *because it's the right thing to do*. Some more recent philosophers find the Kantian view too demanding, and propose the new view as a more lenient and reasonable alternative. Nomy Arpaly (2002) and Julia Markovits (2010) both defend versions of the new view along these lines. Arpaly and Markovits defend a stronger version of the new view than that stated above, as they hold that being motivated by right-making features is necessary for moral worth, as well as sufficient. But I will focus on the sufficiency claim in this paper.

Paulina Sliwa (2016) defends a version of the Kantian view. Her view is also stronger than that stated above; she holds that an act has moral worth iff its agent (a) is motivated to do the right thing by the fact that it is right and (b) knows what the right thing to do is. I do not accept condition (b), for reasons that I will mention in §6. The part of Sliwa's view with which I agree, and that I defend, is the necessity claim above.

On the cinematographic interpretation that I will assume throughout this paper, Finn from Star Wars cares explicitly about the fact that helping Poe to escape is morally right. He has begun to recognize the atrocity of the actions of his comrades and commanding officers, and he wants to break the mold – to disobey orders and choose instead to do what's right, as a small act of rebellion against the First Order's evil regime. As he says, he helps Poe to escape *because it's the right thing to do*.[1] This is the kind of motivation that defenders of the new view denigrate. Building on Michael Smith's (1994, p.75) charge of "moral fetishism", and on Bernard Williams' (1981, p.18) "one thought too many" objection, they suggest that there is something objectionable about the kind of explicitly moral motivation that Finn from Star Wars exhibits. Markovits, for instance, suggests that someone with this motivation is "cold", and is not "a morally attractive person" (2010, p.204).

On the literary interpretation favored by defenders of the new view, Huckleberry Finn is not motivated to help Jim to escape by the fact that doing so is morally right. On the contrary, Huck has *no idea* that what he is doing is right. That is because he has unreflectively absorbed the racist ideology of his contemporaries. According to this ideology, helping a slave to escape constitutes stealing, and is seriously morally wrong. This is why Huck Finn has become a sort of poster child for the new view of moral worth. Defenders of this view cite his example often (e.g. Arpaly 2002, pp.228-31; Arpaly 2003, pp.9-10, 75-78, 92-93, 99-100, 138-39; Markovits 2010, pp.208, 209, 215, 223, 242; Arpaly and Schroeder 2013, pp.178-79, Arpaly 2014, p.63). Their thought is that, since Huck's helping Jim to escape is intuitively a morally worthy act, the case shows that an action can have moral worth even if its agent does not do the right thing because it is right. This case has thus become the go-to counterexample to the Kantian view.

[1] This stipulation may bother avid *Star Wars* fans, who will recall that, as the dialogue progresses, it is suggested that Finn is helping Poe also – or perhaps even solely – because he "needs a pilot" to facilitate his own escape. But Finn wants to escape precisely because his conscience tells him that it is wrong to be complicit in the First Order's evil regime. So I would still construe this as a course of action motivated by the thought that it is morally right.

The literary interpretation favored by defenders of the new view also emphasizes that what motivates Huck to help Jim is the very feature that, in fact, makes this the right thing to do. Arpaly writes that "to the extent that Huckleberry is reluctant to turn Jim in because of Jim's personhood, he *is* acting for morally significant reasons" (p.230, emphasis original). Markovits writes similarly that "he is motivated at least in part by his recognition of Jim's value as a fellow human being – that is, by facts that morally justify his choice" (p.208).

These specifications of the feature that makes Huck's action morally right are conspicuously vague – perhaps deliberately so, to avoid taking too firm of a stand on which first-order moral theory is true. The vagueness will become relevant in §2.2; for now, I simply note that defenders of the new view invite us to assume that Huck Finn is motivated by a right-making feature of his act.

So, according to the Kantian view, Finn from Star Wars performs an action with full moral worth, whereas Huckleberry Finn does not. According to the new view, things are the other way around: Huckleberry Finn performs an action with full moral worth, whereas Finn from Star Wars does not.

Here's where I come in. In this paper I argue against the new view of moral worth, and I defend a version of the Kantian view. I will argue that defenders of the new view are hoisted on their own petard: if Huck really has *no idea whatsoever* that his act is morally right, then his is a case of someone merely *accidentally* doing the right thing. All parties to the historical and contemporary dispute about moral worth agree that an action lacks moral worth if it is a case of someone's merely accidentally doing the right thing. So this means that Huck's action lacks moral worth. So, this case is easy for the Kantian view to accommodate after all: since it is not a case of an action with moral worth, it is no counterexample to the Kantian view.

I begin (in §2.1) by noting that, while there is considerable unclarity as to the nature of moral worth in the existing literature, all parties agree that an action lacks moral worth if it is a case of someone's merely accidentally doing the right thing. I then argue (in §2.2) that the new view cannot adequately account for the phenomenon of accidentally doing the right thing, and that some general reflections on the nature of deliberate action show that the example of Huck Finn – the main example used to support the view – in fact *is* a case of someone accidentally doing the right thing, and thus not of an action with moral worth. I go on to suggest that the new view's plausibility rests on an elision of some important differences between different types of praiseworthiness (§3). Lastly, I offer the beginnings of a defense of one version of the Kantian view by showing how it avoids the problems raised for the new view in this paper (§4). On

my view, all puzzles surrounding the concept of moral worth are just instances of general puzzles about what it is to do something deliberately.

2. Main argument

Here is my argument for the conclusion that Huckleberry Finn's helping Jim to escape lacks moral worth:

1. An action lacks moral worth if it is a case of someone's accidentally doing the right thing.
2. For all types of acts *A*, someone accidentally *A*s if she has no idea that she is performing an act of type *A* when she does so.
3. When Huckleberry Finn helps Jim to escape, he has no idea that doing so is morally right.
4. Someone accidentally does the right thing if she has no idea that she is performing a morally right act when she does so. *(2)*
5. When Huckleberry Finn helps Jim to escape, he accidentally does the right thing. *(3,4)*
6. Huckleberry Finn's helping Jim to escape lacks moral worth. *(1,5)*

This argument is valid. So its success turns on the truth of its three premises. I will defend each in turn.

2.1. Defense of P1

Premise 1 says that an action lacks moral worth if it is a case of someone's accidentally doing the right thing. This is one of the only claims about the nature of moral worth that enjoys anything like widespread consensus across the historical and contemporary literatures on the topic. So we can use this claim to settle disputes about the nature of moral worth in terms that all parties should be able to accept.

"Moral worth" is not an ordinary language term. It originates in English-language translations of Kant's remarks on "moralischen Werth" in the *Groundwork* (1998) and subsequent discussions of Kant's ideas. But it is surprisingly difficult, given this provenance of the concept, to say precisely what moral worth *is*. We do not have a clear definition of the term, summarily recounted by all who employ it. Some philosophers invoke the concept of moral worth without ever saying what it is. And what little is said about the nature of moral worth is not always illuminating.

Here is what we know. Kant introduced the idea of moral worth to distinguish among morally right actions. There are those that are *merely* morally right, and those that have "true moral worth" (G 4:398). Kant thought that the difference has something to do with the agent's motivation for acting; famously, he argues that a morally right act lacks moral worth if it is performed out of self-interest or sympathy for a person in need, and that a morally right act possesses moral worth if it is performed out of a sense of duty.

In the preface to the *Groundwork*, Kant says that acts motivated by immoral aims lack moral worth because these motivations' connection to the moral law is "only very contingent and precarious" (G 4:390). This suggests something about what he thinks the difference is between worth-conferring motivations and non-worth-conferring motivations: it suggests that Kant thinks that worth-conferring motivations bear a connection to the act's rightness that is not "precarious". Barbara Herman takes this line, suggesting that Kant valorizes actions performed out of a sense of duty because this motivation makes acts' moral rightness "the nonaccidental effect of the agent's concern" (1989, p.6). For present purposes I will assume that Kant thought roughly this – I will not take up the exegetical task of working out the details of his view.

Defenders of the new view suggest that acts that are not performed out of a sense of duty may nonetheless have moral worth. In clarifying their disagreement with Kant, these authors offer glosses on the concept of moral worth. But some of these glosses are unhelpful. For example, Arpaly describes the moral worth of an action as "the extent to which the action speaks well of the agent" (2002, p.224), and Markovits says that "morally worthy actions are ones that reflect well on the moral character of the person who performs them" (2010, p.203). These glosses cannot be right. All manner of actions may "speak well of the agent", or "reflect well on [her] character", in that they provide evidence that she has good character. An action need not even be morally right in order to speak well of the agent in this way. For example, imagine a religious group whose members are all extremely virtuous, and who have adopted the convention of saying "Sneezarooney!" after sneezing. Saying "Sneezarooney!" after sneezing speaks well of an agent in this context, as it provides good evidence that she is extremely virtuous. But it is not morally required. And the concept of moral worth, as originally introduced by Kant, is supposed to pick out a property of a proper subset of the morally right actions. So moral worth cannot simply be a matter of an act's speaking well of the agent.

In response to this objection, one might suggest that an act's having moral worth is a matter of its being *both* (a) right *and* (b) performed out of a good motivation. Here is Sliwa (2016, p.1):

> Whether an action is morally praiseworthy depends not just on whether it conforms to the correct normative theory (whatever it is). It needs to be motivated in the right way. An account of moral worth aims to identify what such good motivations consist in.

But in interpreting condition (b) here, we should tread carefully. To say that *any* kind of good motivation leading to the performance of a right act confers moral worth on the act is too strong, and cannot be what Kant had in mind. (Nor is it what Sliwa has in mind – on which see below.) Kant says that benevolent inclinations are praiseworthy, but still do not confer moral worth on actions (G 4:398). So, he explicitly rejects the view that morally worthy actions are those that are both morally right and performed out of a good or praiseworthy motivation.

Moreover, conditions (a) and (b) can be jointly met by actions whose rightness still seems "precarious" in the way that bothered Kant when he was worried about immoral aims. Consider:

> PROMISE-KEEPING: You tell me that you're playing a gig in our local coffee shop at 6pm on Wednesday, and I promise that I'll be there. By the time 6pm Wednesday comes around, I have forgotten all about my promise. But I do want coffee at that time, and I recall that the local coffee shop donates 80% of its profits to charity. This appeals to my desire to be a socially responsible consumer whose purchasing choices contribute to just redistribution of global wealth. So, I go to the coffee shop at 6pm on Wednesday. As I enter and see you strumming away, I realize – with a sigh of relief! – that I have *accidentally* kept my promise.

In PROMISE-KEEPING, I am morally required to go to the local coffee shop at 6pm on Wednesday, since this is what I promised to do. Moreover, the motivation to contribute to just redistribution of global wealth is a good motivation. And this, coupled with my (morally neutral) desire to get coffee at 6pm on Wednesday, motivates me to go to the local coffee shop at 6pm on Wednesday. So, my going to the local coffee shop at 6pm on Wednesday meets conditions (a) and (b) as stated. Yet it still seems as though it is an *accident* that I did the right thing in this case – in exactly the way in which it is an accident that

someone acting on selfish motives does the right thing, if she does. My motivation in this case, though independently praiseworthy, is still only precariously connected to the rightness of my act; contributing to just redistribution of the world's wealth makes my act morally good to do, but what makes it morally *required* is something else (the promise) that does not figure in my motivation at all.[2]

Examples of this form show that conditions (a) and (b) as stated are not jointly sufficient for moral worth – at least, not in the sense that Kant originally had in mind. We might think that these conditions jointly identify something interesting, and stipulate that we use the term "moral worth" to refer to it. But this would not be engaging substantively in a literature borne out of critical engagement with Kant. This would be taking a term from the Kantian secondary literature and unhelpfully using it to refer to something else.

The PROMISE-KEEPING example also highlights a problem with a final recent gloss on the concept of moral worth. Arpaly says, "I shall speak interchangeably of a *morally praiseworthy action* and an *action which has positive moral worth*" (p.224, emphases in original). This is unfortunate, as those phrases are definitely not synonymous. There are many ways for an action to be morally praiseworthy, which we should tease apart and keep apart. In PROMISE-KEEPING, my action is morally praiseworthy, since it embodies a praiseworthy decision to contribute to just redistribution of global wealth. But this is still a case in which the connection between my motivation and the rightness of my act is precarious, and thus in which my action lacks moral worth. So, an action's being morally praiseworthy in *some* way and its possessing moral worth are not the same thing. Rather, moral worth has to do with a *particular* way in which actions can be praiseworthy. (I discuss this a great deal further in §§3-4.)

At this point it may be tempting to abandon hope of identifying an account of the nature of moral worth that is accepted by everyone in the literature. There may be no such thing. This would cast some doubt on the usefulness of the literature.

[2] Here is a recipe for creating counterexamples of this form: take a property of acts that makes them good to do, but not morally required, and take another property of acts that makes them morally required. Imagine an act that has both properties. Then imagine an agent who has no idea about the property that makes the act required, but is nonetheless motivated to perform it by the property that makes it good to do. Voilà! You have a case in which a good motivation leads someone to perform the morally right act, but is only precariously connected to the act's rightness.

I still have hope. This is because I think we can make considerable philosophical progress if we concentrate on one central component of the concept of moral worth, which historical and contemporary authors all accept, and which I have already begun to employ here. I propose that we focus on the idea that, for an action to have moral worth, it must not be a case of someone's merely *accidentally doing the right thing*.

All parties in the contemporary literature accept this idea. In the paragraphs immediately following the quotation above, Sliwa clarifies that she thinks that not just *any* old good motivation confers moral worth on a right act, but only those that prevent the act's rightness from being "contingent and precarious" in the way that bothered Kant (2016, p.2). Sliwa also says that "a central feature of morally worthy actions is that they are not merely accidentally right" (p.6), and that "abandoning the thought that morally worthy actions are non-accidentally right [would be] too high a price to pay" (p.8). Defenders of the new view agree. For example, Markovits writes that Kant's view "gained what attraction it held from the plausibility of the thought that morally worthy actions don't just *happen* to conform to the moral law – as a matter of mere accident" (2010, p.206, emphasis original), and that "[a] plausible account of moral worth… should explain why and how, in the case of morally worthy actions, the connection between the agent's motivations and the act's rightness was not merely accidental" (p.241). She argues that the new view provides just as good an explanation of this non-accidental connection as the Kantian view. (I will discuss her argument in §2.2.) Similarly, Arpaly describes the verdict on Huckleberry Finn that she opposes as the view that he is "a bad boy who has accidentally done something good" (2002, p.230), or "a racist boy who has accidentally done something good" (p.229). She presents herself as denying this in saying that Huck's action has moral worth. So Arpaly accepts that, for an action to have moral worth, it must not be a case of someone's accidentally doing the right (or good) thing. In short, there is clearly some consensus on this point.[3]

There is similar consensus in the Kantian secondary literature. For instance, Marcia Baron writes that "what matters [for moral worth] is that the action is in accord with duty and *it is no accident that it is*" (1995, p.131, emphasis original). And Philip Stratton-Lake, quoting Baron, explains that "the key point

[3] Indeed, the contemporary literature proceeds partly by way of a discussion of which *types* of accidentality limit an action's moral worth. For instance, Arpaly and Markovits disagree about whether the contingency of an agent's being motivated by the right-making features limits her action's moral worth; see Arpaly's remarks on "fair-weather" and "capricious" philanthropists (2002, pp.235-236), and Markovits' "fanatical dog-lover" example (2010, p.210).

about the moral worth of [acting from duty] is that if one does the right act, '*it will be no accident that it is*' right" (2000, p.56, emphasis original), going on to discuss at length what it is for an act's motive to render it non-accidentally right. As we have seen, Barbara Herman also writes that worth-conferring motivations are those that make an act's rightness the "nonaccidental effect of the agent's concern". Indeed, I think that this gloss is the most recognizably Kantian of those that I have canvassed in this section; it seems closer than any other gloss to reflecting Kant's worry about the "precariousness" of agents' acting rightly when moved by immoral aims. The idea that an action lacks moral worth if it is a case of someone's accidentally doing the right thing thus captures what originally bothered Kant when he argued that some right actions lack moral worth.

This gives us only a necessary condition on moral worth, rather than a full analysis. But if we are looking for a central component of the concept accepted by all parties, it may be as good as we can get.

Moreover, this condition can do important philosophical work. We have already seen that this condition shows that moral worth requires more than being moved to do the right thing by a good motivation. I think that we can do better still: I think that this condition shows that the new view is false. Showing this is my task for the next two sections.

2.2. *Defense of P2*

We have seen that there is consensus on the idea that an action lacks moral worth if it is a case of someone's accidentally doing the right thing. But this consensus masks a deeper disagreement. The disagreement is about what it *is* to accidentally do the right thing. Defenders of the new view and the Kantian view assume different general accounts of what it is to do something accidentally. So, we can make some progress in adjudicating the dispute between these views by examining the plausibility of their respective assumptions about what it is to do something accidentally. I will argue that the assumptions underpinning the new view are much less plausible than those underpinning the Kantian view.

Defenders of the new view accept that an action lacks moral worth if it is an instance of someone's merely accidentally doing the right thing. But they hold that someone does *not* accidentally do the right thing if it is the right-making features of the act that motivate them to perform it. When Arpaly denies that Huck is

"a bad boy who has accidentally done something good" (*op. cit.*), her grounds for doing so are that he was motivated to help Jim by the feature of this act that makes it morally good. Markovits agrees, saying that "[a]ctions motivated by right-making reasons… are not merely… accidentally right. If I am motivated by right-making reasons, it is no coincidence that my motive issues in the right action" (2010., p.211).

According to the new view, not just any old motive that issues in morally right action is worth-conferring. Someone who performed the morally right act for selfish reasons would not thereby perform an act with moral worth. Having a motive that *reliably* issues in right action is not enough, either; selfish motives are not worth-conferring even if they reliably lead the agent to act rightly (see Markovits 2010, p.211, n.23). Rather, when Markovits says that it is "no coincidence" that I act rightly if I am motivated by an act's right-making features, she calls our attention to the *metaphysical relationship* between the features of the act that motivate me and its moral rightness. On this view, it is my being motivated by the features that *make* my act right that renders its rightness non-accidental in the manner required for moral worth. On this view, then, the worth-conferring motivations are those that have as their objects the features of acts that bear a certain metaphysical relationship – the "makes it the case" relationship, however this is to be understood – to moral rightness.[4] Non-accidentally doing the right thing amounts to being moved by these features.

Generalizing, we can see the sense of the terms "accident" and "accidental" implicit in this view. On this view, for someone to non-accidentally perform an act of type *A* it is sufficient that (a) she is motivated to perform it by the fact that it is of type *B* and (b) as a matter of metaphysical fact, the agent's performing an act of type *B* makes it the case that she performs an act of type *A*. We are supposed to think that it is "no accident" or "no coincidence" that someone *A*s when, given this metaphysical relationship between the feature that motivates her and the fact that she performs an act of type *A*, it is no *surprise* that she *A*s.

I don't think this is what ordinary speakers of English mean by the terms "accident" and "accidental".

[4] There is some risk of misunderstanding Markovits' account on this point, since Markovits holds that moral reasons are subjective: they are facts that provide evidence about what it would be best to do (e.g. 2010, p.219). Nonetheless, Markovits is clear that she takes such facts to *make* actions right – she holds that the makes-it-the-case relation obtains between the subjective moral reasons that she is interested in and the moral rightness of acts. Thanks to an anonymous reviewer for this journal for encouraging me to clarify this point.

Recall the PROMISE-KEEPING example from §2. In this case, I am motivated to go to our local coffee shop at 6pm on Wednesday. Since going to the coffee shop at 6pm on Wednesday is precisely what I had promised to do, my doing so makes it the case that I keep my promise – it *constitutes* keeping my promise. So in this case I am motivated by the very feature of my act (going to the coffee shop at 6pm on Wednesday) that makes it the case that I perform an act of a further good type (promise-keeping). Yet it still seems as though I *accidentally* perform an act of this further type. Since I have forgotten all about my promise, and thus am entirely unaware of the metaphysical relationship between the feature of the act that motivates me and its being an instance of promise-keeping, I *accidentally* keep my promise.

Here are three more examples, two of them drawn from the philosophical literature on luck and accidents, and one from the philosophical literature on know-how:

> BURIED TREASURE[5]: Vincent wants to plant a rosebush in honor of his dead mother. What he doesn't know is that the one spot on his island that is suitable for growing roses is also the spot where buried treasure lurks just beneath the ground (the pirate who buried the treasure was also fond of roses). So, when Vincent unearths the treasure, he can't believe his luck; how cool to *accidentally* discover buried treasure!
>
> ACCIDENTAL SLAYER[6]: Emilia has been running from vampires all night. Exhausted and desperate to escape, she runs out into an open field at what is, unbeknownst to her, the exact time that the sun's rays peek over the horizon, turning the vampires into dust. Emilia is overcome with relief when she turns around and sees that she has accidentally lured the vampires to their death. (The author of this example names it "Accidental Slayer".)
>
> SEMAPHORE DANCER[7]: A dancer performs a new piece known only as "Improvisation No. 14". A stunned communications expert in the audience notices that this dance is a perfect semaphore rendition of Gray's *Elegy*. But the dancer has no idea about this; she has heard of semaphore but does not know the language, and has heard of Gray's *Elegy* but does

[5] This example is adapted from Lackey (2008).
[6] This example is adapted from Riggs (2014).
[7] This example is adapted from Carr (1979).

not know the poem. The dancer accidentally performs a semaphore rendition of Gray's *Elegy*.

The metaphysical sense of the terms "accident" and "accidental" yields counterintuitive claims about examples like BURIED TREASURE, ACCIDENTAL SLAYER, and SEMAPHORE DANCER. The agents in these cases are each motivated by a feature that makes it the case that they perform an act of a certain type. Yet it still seems natural to say that they *accidentally* perform acts of these types. Vincent is motivated to dig in a certain spot, and his digging in this spot makes it the case that he unearths buried treasure. Yet he still *accidentally* unearths buried treasure. Emilia is motivated to run out into the open field, and this constitutes luring the vampires to their death. Yet she *accidentally* lures the vampires to their death. The dancer is motivated to perform a certain sequence of bodily movements, and this sequence of movements just *is* a semaphore rendition of Gray's *Elegy*. Yet she still *accidentally* performs a semaphore rendition of Gray's *Elegy*.

Why is it that, in cases like PROMISE-KEEPING, BURIED TREASURE, ACCIDENTAL SLAYER, and SEMAPHORE DANCER, it is natural to say that the agent *accidentally* performs an act of a certain type, though they are each motivated by the feature that makes it the case that they perform an act of the relevant type?

Consider the dancer. She is motivated to perform a certain sequence of movements. And this sequence is, in fact, semaphore-rendition-of-Gray's-*Elegy*-making. Yet she accidentally performs a semaphore rendition of Gray's *Elegy*. This is because she does not *mean* to perform a semaphore rendition of Gray's Elegy, nor does she believe that she is doing so, nor does she have even a vague inkling that her dance may constitute a semaphore rendition of Gray's *Elegy*. Were she to learn that she had performed a semaphore rendition of Gray's *Elegy*, she would be astonished. In short, the dancer has *no idea whatsoever* that her dance is a semaphore rendition of Gray's *Elegy*. This is what makes us inclined to say that she accidentally performs a semaphore rendition of Gray's *Elegy*, notwithstanding the metaphysical relationship between her dance and the language of semaphore. Parallel remarks apply to the other cases.

These are not isolated examples. On the contrary, it is easy to come up with cases like this. Here is a recipe: construct a scenario in which (a) an agent is motivated to perform an act by the fact that it is of type *B*, and (b) as a matter of metaphysical fact, the agent's performing an act of type *B* makes it the case that she performs an act of type *A*, but (c) the agent is wholly unaware of (b). Voilà! You have a scenario

in which it seems natural to say that the agent accidentally *A*s, though she is motivated to perform her act by its *A*-making feature. Indeed, I suspect that, the more emphasis we place on the fact that she has *no idea whatsoever* that the feature of the act that is motivating her is *A*-making, the more it will seem that she does an *A* thing by accident. This is so notwithstanding the fact that, in the new view's sense of "accident", it is "no accident" that the agent performs an act of type *A* just as long as (a) and (b) hold.

At this point, defenders of the new view might object. They may note that, in each of my examples, the feature of the act by which the agent is motivated makes it the case that the agent performs an act of a further accidental-seeming type only given some important background conditions: the fact that the treasure is buried in Vincent's chosen spot, the fact that the sun is about to rise over Emilia's field, the facts about the language of semaphore and the content of Gray's *Elegy*, and the fact that I promised to go to the coffee shop at 6pm. An objector may claim that this is a crucial disanalogy, and that the metaphysical relationship between right-making features and rightness is less dependent on background conditions than the metaphysical relationships in my cases.

In response, I agree that background conditions play an important role in my cases. But I maintain that the metaphysical relationship between right-making features and rightness is no less dependent on background conditions. Consider Huck Finn again. It is simply false to say that background conditions play less of a role in his case than in my cases. Huck is not motivated by something that *necessitates* the rightness of his act. Rather, he is motivated by something that can make his act right only given certain crucial background conditions. If Jim were a serial killer on the run, rather than a fugitive slave, then facts about his personhood would not make it morally right to help him to escape from the authorities. So, the defender of the new view must concede that her view is already about cases in which an agent's performing an act of one type (protecting a person) makes it the case that she performs an act of another type (morally right act) only given important background conditions. Once she has conceded this point, my examples are fair game.

The defender of the new view might insist that, if we spell out the content of Huck's motivation in full detail, we will find that he *is* motivated by something that necessitates the rightness of his act, all by itself, requiring no background conditions. I seriously doubt this. To get a feature that necessitates moral rightness, we would need to specify the feature in an inordinate amount of detail; we would need to build the absence of any circumstances that would create a counterexample into our specification of the

feature itself. For instance, we would have to say that the feature of the act that motivates Huck is not just that it helps a person but rather that it helps a person *who is trying to do something that is itself morally valuable, and who is not hurting anyone else in the process, and who is not disrupting any social institutions besides those that are harmful and should be disrupted*, et cetera. Without these qualifications, we will not be specifying a feature that necessitates the moral rightness of the act, but rather a feature that is compatible with an act's being morally wrong under some circumstances. So, to get a feature that necessitates the rightness of the act, the defender of the new view needs all these qualifications. But, with the qualifications, this feature is simply too complicated to be the object of Huck's motivation. Huck's motivation – like those of many moral agents – is far too inchoate and rudimentary to have such a complex property as its object. So, he is not motivated by a feature of his act that necessitates its moral rightness. Rather, as I have assumed, he is motivated by a feature that makes his act right only given some important background conditions.

Returning to the main argument, I think that we now have good grounds to accept the following claim:

> **CLAIM**: For all types of acts *A*, someone accidentally *A*s if she has no idea that she is performing an act of type *A* when she does so.

We have seen that **CLAIM** holds even of cases in which an agent is motivated by the very feature that makes it the case that she performs an act of type *A* (in light of some background conditions). So long as she has no idea whatsoever that this metaphysical relationship obtains between the feature of the act that motivates her and its being an act of type *A*, she still accidentally *A*s.

CLAIM is premise 2 from my argument above.

This concludes my defense of premise 2.

2.3. Defense of P3

This just leaves premise 3: when Huckleberry Finn helps Jim to escape, he has no idea that doing so is morally right.

Defenders of the new view of moral worth are explicitly committed to this claim. Indeed, it is crucial for them that this is true, as otherwise the example of Huckleberry Finn cannot do the philosophical work to which they have tried to put it.

The actual character portrayed in Twain's text is not a great counterexample to the Kantian view. A natural reading of the text is to say that Huck has an inchoate grasp of the moral rightness of this act, or that he believes that it is morally right "at some level", or something along these lines, and that this is why he does it. But if any such interpretation is correct, then the case is no counterexample to the Kantian view. On any such interpretation, the case would call only for a modification of the Kantian view to allow for the evident fact that it is possible for someone to be motivated by something that she may not consciously avow, but grasps at the subpersonal level. This is a modification that the Kantian view must undergo anyway on grounds of phenomenological plausibility and fit with contemporary psychology on motivation.

To provide a clear counterexample, then, defenders of the new view must employ a certain interpretation of Twain's text. Here are some representative quotations (emphases original):

> As the familiar case of Mark Twain's Huckleberry Finn shows, an act can have moral worth even if it is performed in the belief that it is *wrong*. (Markovits 2010, p.208)

> [M]y point is not simply that Huckleberry does not have the belief that his action is moral on his mind when he acts. He does not have the belief that what he does is right *anywhere* in his head. (Arpaly 2002, p.229)

This is an interpretation on which Huck fully believes that his act is wrong, where this precludes his also believing that it is right. On this interpretation, he has no subpersonal grasp of his act's rightness; he has no belief in its rightness "anywhere in his head". This phrase is not just a rhetorical flourish. For the example of Huckleberry Finn to be a clear counterexample to the Kantian view, it is crucial for defenders of the new view to emphasize – as, indeed, they do – that Huck has *no idea whatsoever* that his act is right.

But this means that Huck's position with respect to the rightness of his act is like the dancer's position with respect to her dance's being a perfect semaphore rendition of Gray's *Elegy*. He does not mean to act

rightly, nor does he believe, or even have a vague inkling, that his activity might constitute doing what's morally right. Were he to learn that his act is morally right, he would be astonished.

Indeed, if anything, Huck Finn is doing *worse* than the dancer. She presumably has not even considered the possibility that her dance is a perfect semaphore rendition of Gray's *Elegy*. But at least she has not actively considered this possibility and explicitly ruled it out. Huck, by contrast, has considered the moral status of his act, and is fully convinced that it is wrong. So he does not simply lack a belief about his act's rightness: he has a false belief. To make my examples more closely analogous to Huck's case, then, we would have to stipulate that the agents are explicitly convinced that they are *not* performing acts of the relevant types. But this would only make it seem clearer that they accidentally perform acts of the relevant types. For instance, if Vincent is fully convinced that his island is utterly devoid of treasure, and thus that by digging in his chosen spot he will definitely not unearth buried treasure, then it seems particularly clear that when he in fact unearths buried treasure, he does so only accidentally. Parallel remarks apply to the other cases.

If we are going to say such things about these other agents, then we should say them about Huck Finn, too. To wit: if Huck Finn is *fully convinced* that his helping Jim to escape is morally wrong, and he has *no idea* that it is in fact morally right – if he does not have the belief that his act is right "anywhere in his head" – then, in helping Jim to escape, he accidentally does the right thing.

But we are granting to defenders of the new view the assumption that Huck is motivated by what is, in fact, the right-making feature of his act (in light of some background conditions). And we are also taking for granted the shared assumption that an action lacks moral worth if it is a case of someone's accidentally doing the right thing. So, this argument shows that being motivated to do the right thing by the feature that makes it right is insufficient for moral worth. In other words, it shows that the new view is false.

3. Where did the new view go wrong?

I have argued that the new view of moral worth – the view that an action has moral worth if its agent was motivated to do the right thing by the features that make it right – is false. This view cannot be squared

with the idea that an action lacks moral worth if it is a case of someone's accidentally doing the right thing. The construal of the terms "accident" and "accidental" that we must accept to render this idea consistent with the new view is an implausible construal that flies in the face of ordinary intuitions about the extension of these terms.

But some brilliant philosophers have defended the new view. So what went wrong?

As discussed in §2.1, there is a notable lack of clarity in the existing literature about what moral worth *is*. I suspect that this lack of clarity has led us astray. Recall the glosses on the concept of moral worth that I criticized earlier: they were put in terms of an act's "reflecting well" on an agent's character, or "speaking well" of her, or of her being led to perform the right act by a good motivation. I argued earlier that none of these glosses captures what Kant had in mind when he complained of the "precarious" connection between an agent's motivation and her act's rightness that characterizes a right act without moral worth. Nonetheless, I think, the glosses are getting at something.

What they are getting at is the close connection between moral worth and *praiseworthiness*. Not just any old kind of praiseworthiness confers moral worth on actions – this was one of the lessons of §2.1. But there is still an important connection between moral worth and a particular kind of praiseworthiness: when someone performs an act with moral worth, she is praiseworthy *for acting rightly*, whereas if someone's act is morally right but lacks moral worth, then she is not praiseworthy for acting rightly. (One might wonder why there is this connection between two good things about an agent and her act. I offer my explanation of the connection in §4.)

This connection between performing an act with moral worth and being praiseworthy for acting rightly is widely presupposed in the contemporary literature, both by defenders of the new view and defenders of the Kantian view. For example, Arpaly spells out her account of moral worth without actually using the phrase "moral worth", instead writing about what it takes to be praiseworthy for doing the right thing: her account is that "for an agent to be morally praiseworthy for doing the right thing is for her to have done the right thing for the relevant moral reasons" (2002, p.226). To present this claim as an account of moral worth is to presuppose that there is a close connection between an act's having moral worth and an agent's being morally praiseworthy for doing the right thing. Similarly, Sliwa notes that an agent "seems

praiseworthy for doing the right thing" under precisely the circumstances in which "it's not just a fluke that [the agent] gets it right" – i.e., in which she performs an act with moral worth (2015, p.19).

I think that the new view goes astray here by eliding an important distinction between different kinds of praiseworthiness. Defenders of this view trade heavily on the popular intuition that there seems something praiseworthy about Huck Finn. This is accurate. There seems *something* praiseworthy about him. But notice that the popular intuition about Huck is not specifically that he performs an action with full moral worth. The term "moral worth" is not an everyday term; it is a philosophers' term of art. Ordinary people's positive reactions to Huck Finn suggest that we take there to be *something* good about him, but they leave open what exactly this good thing is. The literature on different types of praiseworthiness is still young, so it is worth carefully teasing and then keeping apart the many different species of this genus, bearing in mind that an agent may enjoy some but not all of them.

I think that the new view elides the distinction between two kinds of praiseworthiness: being praiseworthy *for having a good character trait* and being praiseworthy *for performing a good type of act*. It also conflates being praiseworthy for performing an act of type *A* with being praiseworthy for performing an act of type *B*, where the act's being of type *B* metaphysically constitutes its being of type *A*. I'll now explain what I mean by this.

When someone is praiseworthy for acting rightly, she is praiseworthy in virtue of having performed a good type of act: a morally right act. This is a normal sort of thing. We are often praiseworthy in virtue of our performing acts of good types. For example, someone can be praiseworthy for doing stuff that constitutes keeping a promise, or helping her sister, or buying the groceries, or making a pun. In such cases she is praiseworthy *for performing a certain type of act*. Likewise, when someone non-accidentally does the right thing in the manner characteristic of moral worth, she is praiseworthy for doing stuff that constitutes acting morally rightly. This is also a way of being praiseworthy for performing a certain type of act.

It is important to distinguish *de re* and *de dicto* readings of the above. There are two ways to hear the claim that someone is praiseworthy for doing stuff that constitutes keeping a promise. On one reading, there is some stuff that the agent is praiseworthy for doing (*de re*), and this stuff constitutes keeping a promise. On the other reading, what the agent is praiseworthy for is *doing stuff that constitutes keeping a promise* (*de*

dicto). These come apart. For example, in PROMISE-KEEPING, there is some stuff I do: financially supporting a local business that donates an overwhelming portion of its profits to charity. I am praiseworthy for doing this, since it is benevolent. And, in context, this activity constitutes keeping a promise. But that does not mean that I am praiseworthy *for doing stuff that constitutes keeping a promise* (*de dicto*). I am not praiseworthy for keeping my promise, since I have no idea that I am keeping it, and thus I do so only accidentally.

I think that parallel remarks apply to Huck Finn. There is something he does: protecting a person. He is praiseworthy for doing this, since it is benevolent. And, in context, this activity constitutes acting rightly. But this does not mean that Huck is praiseworthy *for doing stuff that constitutes acting rightly* (*de dicto*). Huck is not praiseworthy for acting rightly, since he has no idea that he is doing so, and thus he does the right thing only accidentally.

As well as being praiseworthy for performing acts of certain types, we can be praiseworthy for character traits. For example, the desire to be a socially responsible consumer is a praiseworthy character trait. So in PROMISE-KEEPING I am praiseworthy for this character trait. More generally, whenever someone is motivated by a right-making feature – for example, when she cares about making others feel good in the manner characteristic of kindness, or when she cares about distributing burdens and benefits on reasonable grounds in the manner characteristic of fairness – this is a praiseworthy character trait. So being motivated by a right-making feature is a way of being praiseworthy.

If someone is praiseworthy for a character trait, then she is praiseworthy for having it just as long as she continues to have it, regardless of whether it manifests in her behavior throughout this time. (This is why, when someone praises an agent for her kindness, it is no objection to say "But she's sleeping currently!") Likewise, someone can be praiseworthy for performing a particular act of a certain good type even if she has no stable disposition to perform acts of that type, and no corresponding praiseworthy character trait. (If I help my sister on one occasion, I can be praiseworthy for helping her on this one occasion even if I have no corresponding character trait and I usually do very little to help her.) So praise for having good character traits and praise for performing good types of act can vary independently of one another.

Praiseworthy character traits do sometimes manifest in our action. When this happens, they often lead us to perform acts of at least one good type. In all such cases, the agent is praiseworthy for the character trait

manifested in her act – she is praiseworthy for it just as long as she has it. But she may not be praiseworthy for having performed an act of the relevant good type. The PROMISE-KEEPING example illustrates this again. In this example, I am led by a praiseworthy trait (benevolence) to do something that constitutes keeping a promise, which is a good type of act. But I am not praiseworthy *for* keeping a promise, since I did so only accidentally. At best, I am praiseworthy (for my character trait) *while* keeping a promise, in the way that somebody can be wearing a hat while walking: these are simply two things that are true of me at the same time.

The three cases in §2.2 illustrate this point equally well. Vincent is praiseworthy for wanting to honor his dead mother, and this praiseworthy motivation leads him to perform an act that constitutes unearthing buried treasure. But he is not praiseworthy *for* unearthing buried treasure. Likewise, even if we stipulate that Emilia and the dancer manifest praiseworthy character traits (of some kind) in running into the open field and performing the dance, this does not make them praiseworthy *for luring the vampires to their death* or *for performing a semaphore rendition of Gray's Elegy*. We are not praiseworthy for that which we do accidentally. This holds even if we are praiseworthy for a good character trait that manifests in the activity that constitutes our performing an act of a certain good type. That makes us praiseworthy *while* performing an act of a certain type. But it is not sufficient for being praiseworthy *for* performing an act of a certain type.

This is where I think the new view goes wrong. Huckleberry Finn has a praiseworthy character trait: he cares about Jim. And this leads him to perform an act of a good type: it is morally right. But it does not follow that Huck is praiseworthy for performing an act of this type. And, in fact, he is *not* praiseworthy for performing an act of this type. Huck accidentally does the right thing, and we are not praiseworthy for that which we do accidentally. So, Huck lacks the particular kind of praiseworthiness that is the mark of an act with moral worth. Defenders of the new view think otherwise because they elide the distinction between praise for act-types and praise for character traits; their account of what it is to be praiseworthy *for* acting rightly is in fact just a way of being praiseworthy *while* acting rightly.

More broadly, I think that defenders of the new view take there to be a much closer connection between praise for character traits and praise for act-types than actually obtains. Arpaly thinks that an agent is *more* praiseworthy for acting rightly, "the stronger the moral concern that has led to her action" (2002, p.233). Markovits does not say this, but does say that "morally worthy actions are the building blocks of

virtue – a pattern of performing them makes up the life of a good person" (2010, p.203). I think that neither of these claims is quite correct, though they are both close to something correct. It is true that an agent who acts rightly and has stronger moral concern will be more praiseworthy *overall* than one who acts rightly but has weaker moral concern (under otherwise identical circumstances). But the first agent is not more praiseworthy *for acting rightly* than the other. She is more praiseworthy for her stronger moral concern; it is her *character* that is more praiseworthy. It is also true that morally worthy actions – or, more broadly, actions of good types – are *among* the things for which an agent can be praiseworthy, so their repeated performance contributes cumulatively to an agent's overall praiseworthiness. Such actions are, in this sense, "building blocks of virtue". But they are not *the* building blocks of virtue. There are other things that contribute positively to an agent's overall praiseworthiness: her character traits. And, as I have emphasized, praiseworthy character traits and the performance of praiseworthy actions do not systematically co-vary. Someone who has a praiseworthy character trait is likely to manifest it in her action by performing acts of good types, and is likely to have developed it by practicing performing acts of good types. Similarly, someone who deliberately performs acts of good types must have *some* praiseworthy character traits, since the motivation to perform acts of these good types, however weak, is itself a praiseworthy character trait. But that's as close as the connection gets.

4. Deliberately doing the right thing

I have argued against the new view of moral worth. In this section, I will give the brief beginnings of a defense of one version of the Kantian view. To fully defend this view would take several more papers. But I will explain how this version of the Kantian view avoids the difficulties I have raised for the new view.

Here is my view:

> **MY VIEW**: An act has moral worth just in case it is an instance of someone's *deliberately* doing the right thing.

This is a version of the Kantian view. The Kantian view says that an act has moral worth only if its agent was motivated to do the right thing by the very fact that it is right. I think that the best way to develop this view is to take the performance of an act with moral worth to be a kind of achievement: the

achievement of someone's trying to act rightly and succeeding.[8] On my view, there is a special kind of value in people's deliberately doing the right thing – as when Finn from *Star Wars* helps Poe to escape because it's the right thing to do. Such cases exhibit a kind of achievement that makes them better (in one respect) than cases in which people manage to do the right thing without trying, including cases in which the latter agents are independently praiseworthy in light of their good motivations.

My view offers neat explanations for many of the phenomena discussed above.

To begin at the beginning: it is clear, on my view, why moral worth is a status for which moral rightness is necessary but insufficient. Someone can deliberately do the right thing only if she does the right thing. This is because "deliberately *A*-ing" is a success term; one can *A* deliberately only if one does in fact *A*. (This is a clearer account than that offered by the "speaks well" or "reflects well" glosses from §2, neither of which entails that moral rightness is necessary for moral worth).

It is also quite easy, on my view, to account for the central claim about moral worth that has been the focus of this paper: that an act lacks moral worth if it is a case of someone's accidentally doing the right thing. The terms "deliberate" and "accidental" are antonyms, and the categories to which they refer are logical contraries; if someone does something deliberately, then she does not do it accidentally, and *vice versa*. So someone's accidentally doing the right thing precludes her deliberately doing the right thing. Since an act has moral worth just in case it is an instance of someone's deliberately doing the right thing, someone's accidentally doing the right thing precludes her act's having moral worth.

It is also easy, on my view, to explain the connection between performing an act with moral worth and being praiseworthy for acting rightly (discussed in §3). These states have the same precondition: the agent's deliberately doing the right thing.

This requires some spelling out. In general, we merit praise for performing an act of a certain good type if and only if we do so deliberately, where this contrasts both with what we do accidentally and with what we foresee that we will do but do not intend to do. This is an important lesson to draw from the literature on so-called "Knobe cases" (see Knobe 2003, 2006; Knobe and Pettit 2009). This literature documents a

[8] I think of achievements in roughly the way explicated by Bradford (2015).

pattern whereby experimental subjects describe an agent as "intentionally" bringing about a foreseen side-effect if the effect is bad, but deny that she intentionally brings about a foreseen side-effect if it is good. Knobe's explanation for this asymmetry is that our judgments about the agent's praise- or blameworthiness for bringing about the effect alter our inclination to describe it as intentional. This explanation is based on the observation that Knobe's experimental subjects said that agents deserve a lot of blame for bad side-effects that are foreseen but unintended, but that we deserve very little praise for good side-effects that are foreseen but unintended (Knobe 2006, p.193). I do not think that these people's evaluative intuitions are mistaken or confused on this point. On the contrary, their reactions highlight an asymmetry between praise and blame: we can be blamed for performing an act of a bad type as long as we are aware that our act is of that type, but we merit praise for performing an act of a good type only if we do so deliberately.

If this is correct, it follows that we merit praise for doing the right thing only if we do so deliberately. For example, if Finn from *Star Wars* knew that it was morally right to help Poe to escape, but just did it out of perverse amusement or a desire to be close to people whose names begin with the letter P – so he foresaw that he was acting rightly but did not intend to do so – then he would not merit praise for doing the right thing. Since it is also the case (on my view) that someone performs an act with moral worth iff she deliberately does the right thing, this explains the connection between being praiseworthy for acting rightly and performing an act with moral worth: both require that the agent does the right thing deliberately.

There are some complications here. As we have seen, the accidental and the deliberate are logical contraries, but they not contradictories. Someone can *A* neither deliberately nor accidentally, if she foresees that her act is of type *A* but does not choose to perform it on this basis. Further complications arise from the fact that foresight comes in degrees. There are all manner of doxastic attitudes that someone can take toward the fact that she is performing an act of some good type that are better than having *no idea whatsoever* that she is performing an act of this type. For example, she could have a vague inkling that she is performing an act of this type, or she could have credence 0.2467 that she is doing so. So there are lots of open questions for my view concerning what to say about someone who has one of these intermediate attitudes toward the fact that she is acting rightly.

We can settle some cases based on what I have said so far. Often, when someone takes an intermediate doxastic attitude toward the fact that her act is right, she does not then choose to perform it *on the basis of* its (possible) rightness. In performing the act, she is not *trying* to act rightly. In such cases the act's rightness is foreseen to some degree, but is not intended. On my view, this act determinately lacks moral worth. Since it is not performed on the basis of its (possible) rightness, it is not an instance of someone's deliberately doing the right thing. In such cases, hard questions about the degree of the agent's foresight are happily irrelevant to the question of whether her act has moral worth.

There are other cases in which the details of an agent's doxastic attitude toward the fact that her act is right can make a big difference. These are cases in which the agent chooses to perform the act on the basis of its (possible) rightness, though she is not sure that it is right. Here I think it can be unclear whether she counts as deliberately doing the right thing, and thus unclear whether her act has moral worth. But these cases are just instances of a general puzzle in philosophy of action: it is unclear what doxastic attitude someone must take toward the fact that she is performing an act of a certain type to count as *deliberately* performing an act of this type, when she in fact succeeds in doing so. This puzzle goes back at least as far as Davidson (1971) and Bratman (1984), and the ensuing literature on intention without belief.

I am inclined to be lenient in such cases. I think that someone can deliberately perform an act with a certain property even if she has very little credence that her act has the property when she performs it, provided that performing an act with this property is precisely what she was trying to do all along. This is a point of disagreement between my account and Sliwa's (2016) account. As mentioned above (§1), Sliwa holds that an act has moral worth iff the agent does it because it is right *and knows that it is right*. I think that this knowledge is unnecessary, as I do not think that knowledge of what one is doing is necessary in order to count as doing something deliberately. For example, consider Finn from *Star Wars* again. Does he *deliberately* save Poe from the First Order? Yes. But does he *know* that he is saving Poe, at the time when he does so? I think not. As a trained Stormtrooper, Finn is well aware of the Star Destroyer's technological capacities. And the Destroyer is an extremely powerful ship: powerful enough to destroy the TIE fighter in which Finn and Poe escape (though in the actual movie it only damages the TIE fighter and sends them hurtling down onto Jakku). So Finn's evidence does not warrant his being sufficiently confident of success for him to *know* that he is saving Poe. Nonetheless, when he does succeed, it would seem churlish to deny that he saved Poe deliberately. The lesson to draw is that deliberately *A*-ing requires only minimal foresight that one is in fact *A*-ing, if this is precisely what one

was *trying* to do all along.[9] Analogously, someone can deliberately do the right thing even if she is not at all confident that her act is morally right, if she is trying to act rightly and succeeding.

There are other genres of puzzle case in the vicinity. As is well-recognized in the literature on moral worth, there is a question of what to say about agents who do the right thing because it's the right thing to do, but accept a mistaken moral theory, and thus are mistaken about what the act's right-making features are (cf. e.g. Arpaly 2002, p.227; Sliwa 2016, pp.4-5). In extreme cases, in which a radically mistaken agent takes things to be her act's right-making features that are totally different from its actual right-making features, it is natural to describe her as having hit upon the right act *by accident*. For example, if Finn from *Star Wars* thinks that it is morally right to help Poe to escape just because he once heard that men in leather jackets should never be kept in captivity, then it is natural to describe him as having hit upon the morally right act by accident. There is a related question of what to say about agents who aim to do the right thing and are caused by their aim do the right thing, but via a deviant causal chain (analogous to the climber example in Davidson 1973, pp.153-154). For example, someone could want to act rightly and recognize that ϕ-ing is the right thing to do but be so nervous about this prospect that she goes into convulsions, involuntarily performing the precise sequence of bodily movements constitutive of ϕ-ing.

These cases are puzzling. But, again, these puzzles are not special problems for my view of moral worth. They are general puzzles in the philosophy of action about what it takes to do something deliberately. In general, it is unclear how wrong someone can be about why her activity constitutes *A*-ing in order to count as deliberately *A*-ing. And it is unclear how best to spell out the concept of deliberate action so as to exclude deviant causal chains. Since my view employs the idea of doing something deliberately, it

[9] Those attracted to a certain way of thinking about the terms "accident" and "accidental" may worry about this view. On one way of thinking, popular in the post-Gettier literature in epistemology, non-accidentally *A*-ing requires counterfactual success: the agent must *A* not only in the actual world, but also in a range of nearby possible worlds. Those attracted to this approach may worry that my view does too little to ensure counterfactual success. If someone can *A* deliberately without knowing that she is *A*-ing, then what guarantees that she *A*s in a nearby range of possible worlds? My answer is that there is a degree of counterfactual success built in to the concept of deliberate action. If someone *A*s deliberately, then she wants to *A* and succeeds in *A*-ing by exercising effort and skill; this differentiates *A*-ing deliberately from *A*-ing accidentally, with mere foresight, or as part of a deviant causal chain. But the motivation, effort, and skill constitutive of *A*-ing deliberately together ensure the agent's success in *A*-ing in some nearby worlds. Beyond this, there are no counterfactual guarantees: an agent's success in doing what she tries to do requires a favorable set of surrounding circumstances, which may differ in nearby worlds. But I think that this should not worry us. Whether an agent deserves praise for performing a good act does not depend on the successes of all her counterparts, but just on whether the performance counts as an achievement for *her*, the actual person.

inherits these puzzles. But I am hopeful that the most promising general solutions, whatever they turn out to be, will be applicable here. Indeed, one nice feature of my view of moral worth is that it shows that some key puzzles about this concept are just instances of general puzzles about the nature of deliberate action.

Here is one last perk of my view: the view provides clear and simple answers to two questions that have dominated the contemporary Kantian literature on moral worth. Much has been written on whether moral worth requires that an agent perform an act *only* because it's the right thing to do, as opposed to performing it *both* because it's right *and* for some other reason. Much has also been written on how explicitly an agent must consider the fact that her act is morally right, when choosing to perform it, for her action to have moral worth (For detailed discussion of both questions see e.g. Henson 1979; Herman 1981; Baron 1995, ch.4-5; Stratton-Lake 2000, ch.3-4.) My view offers simple answers to both questions: the action has moral worth just in case the agent counts as deliberately doing the right thing. This answers the first question. It is possible to do something for more than one reason, so it is possible to do something both because it's the right thing to do and for some other reason. On my view, so long as the agent counts as deliberately doing the right thing, her action has moral worth. This also answers the second question: an agent must consider the rightness of her act however explicitly is necessary to count as deliberately doing the right thing. It is perfectly possible to deliberately *A* without thinking furiously of the fact that one is *A*-ing throughout the duration of this activity. One's awareness of the fact that one is *A*-ing can operate at the subpersonal level. On my view, so long as the agent's attitude toward her act's moral rightness is still sufficient for her to count as deliberately doing the right thing, her action still has moral worth.

Thus my preferred version of the Kantian view avoids the difficulties that I have raised for the new view. And, while the view faces puzzles of its own, these are familiar puzzles surrounding the idea of deliberate action, the solving of which can be allocated to philosophers working directly on these puzzles. I hope that this is enough to make the Kantian view seem worth reconsidering.[10]

[10] This paper has been a long time in the making and has benefited from the help of very many people. I am grateful to Nomy Arpaly, Sarah Buss, David Faraci, Scott Hershovitz, Nathan Howard, Maria Lasonen-Aarnio, Rob Long, Ralph Wedgwood, and Brian Weatherson, for helpful comments on earlier drafts. I am also grateful to audiences at the Princeton-Michigan Workshop on Metanormativity 2016, the Great Plains Philosophy Symposium 2016, the USC-UCLA Graduate Conference 2017, the Rocky Mountain Ethics Congress 2017, and at the University of Michigan, the University of Southern California, the University of Notre Dame, Simon Fraser University, Florida State University,

REFERENCES

Arpaly, Nomy (2002). "Moral Worth". *The Journal of Philosophy* 99(5), pp.223-245.

Arpaly, Nomy (2003). *Unprincipled Virtue*. New York, NY: Oxford University Press.

Arpaly, Nomy (2014). "Duty, Desire, and the Good Person: Towards a Non-Aristotelian Account of Virtue". *Philosophical Perspectives* 28, pp.59-74.

Arpaly, Nomy and Schroeder, Timothy (2013). *In Praise of Desire*. New York, NY: Oxford University Press.

Baron, Marcia (1995). *Kantian Ethics Almost Without Apology*. Ithaca, NY: Cornell University Press.

Bennett, Jonathan (1974). "The Conscience of Huckleberry Finn". *Philosophy* 49(188), pp.123-134.

Bradford, Gwen (2015). *Achievement*. Oxford, UK: Oxford University Press.

Bratman, Michael (1984). "Two Faces of Intention". *The Philosophical Review* 93(3), pp.375-405.

Carr, David (1979). "The Logic of Knowing How and Ability". *Mind* 88, pp.394-409.

Davidson, Donald (1971). "Agency". In R. Ausonio Marras, N. Bronaugh and R. W. Binkley, eds., *Agent, Action, and Reason*, pp. 1-37. Toronto, ON: University of Toronto Press.

Davidson, Donald (1973). "Freedom to Act". In T. Honderich, ed., *Essays on Freedom of Action*, pp.137-155.

Henson, Richard (1979). "What Kant Might Have Said: Moral Worth and the Overdetermination of Dutiful Action". *The Philosophical Review* 88, 39-54.

Herman, Barbara (1981). "On the Value of Acting from the Motive of Duty". *The Philosophical Review* 90(3), pp.359-382.

Kant, Immanuel (1785, repr. 1998). *Groundwork of the Metaphysics of Morals*. Trans. Mary Gregor. Cambridge, UK: Cambridge University Press.

Knobe, Joshua (2003). "Intentional Action and Side-Effects in Ordinary Language". *Analysis* 63, pp.160-163.

Knobe, Joshua (2006). "The Concept of an Intentional Action: a Case Study in the Uses of Folk Psychology". *Philosophical Studies* 130, pp.203-231.

McGill University, the University of Miami, the University of Chicago, and the University of California at Santa Barbara. Special thanks go to my advisor, Brian Weatherson, for being the most gracious defender of a false view that I have so far encountered.

Knobe, Joshua, and Pettit, Philip (2009). "The pervasive impact of moral judgment". *Mind and Language* 24, pp.586-604.

Lackey, Jennifer (2008). "What Luck is Not". *Australasion Journal of Philosophy* 86(2), pp.255-267.

Markovits, Julia (2010). "Acting for the Right Reasons". *The Philosophical Review* 119(2), pp.201-242.

Riggs, Wayne (2014). "Luck, Knowledge, and 'Mere' Coincidence". *Metaphilosophy* 45(4-5), pp.627-639.

Sliwa, Paulina (2016). "Moral Worth and Moral Knowledge". *Philosophy and Phenomenological Research* 93(2), pp.393-418.

Smith, Michael (1994). *The Moral Problem*. Oxford: Blackwell.

Stratton-Lake, Philip (2000). *Kant, Duty and Moral Worth*. London: Routledge.

Twain, Mark (1884). *The Adventures of Huckleberry Finn.* London: Chatto & Windus.

Williams, Bernard (1981). "Persons, Character and Morality". In B. Williams, *Moral Luck*, Cambridge University Press, pp.1-18.