The Explanatory Structure of Moral Worth

Harjit Bhogal

Abstract

If an action is merely accidentally, or coincidentally, right then it does not have *moral worth*. This thought drives the central dispute about moral worth – whether worthy actions are motivated by the *fact that the action is right* or by the *right-making features* of an action.

Many find it intuitive that we should be motivated by concrete right-making features rather than abstract facts about rightness. But many others, from Kant onwards, argue that we must be motivated by *rightness itself* in order to avoid doing the right thing merely coincidentally.

I defend a novel account of moral worth, in the spirit of *right-making features* approaches but which avoids such coincidentality worries. A key motivation for the view comes from investigating the concepts of accidentality and coincidence as they apply in various other debates across philosophy.

Keywords: Moral Worth, Coincidences, Explanation

Some actions are morally right. That is, sometimes the action performed matches the facts about rightness – ϕ was performed and ϕ is right.

Of these right actions, only some have moral worth. (I'm assuming, as is common in the literature, that morally worthy actions are objectively morally right.)[1] The politician who does charity work solely for publicity [Sliwa, 2016, p. 393] and the person who acts randomly but luckily don't act worthily. They don't deserve credit for their right actions.

Morally worthy actions, it seems, must be done for the right reasons. There are two main approaches to which reasons are right:

Rightness Itself (RI): Morally worthy actions are motivated by the fact that the action is right

Rightness Making Features (RMF): Morally worthy actions are motivated by the features of the action that make it right[2]

[1] [Sliwa, 2016, section 5] defends this assumption against natural concerns. But I'm open to the possibility that the core view of this paper could be made consistent with the alternative claim – made, for example, by Markovits [2010] and Howard [2021] – that morally worthy actions have to be *subjectively* right. I'll come back to this briefly in footnote 14 but I won't have space to discuss it substantially.

[2] These approaches go by various names. This terminology is from Singh [2020].

These are slogans, not precise formulations, but they point to the central division in the literature. Arpaly [2002] and Markovits [2010], among others, hold RMF views – a worthy action is motivated by the fact that it would reduce suffering, or would keep a promise to your friend, or so on. Others, in the spirit of Kant, [e.g. Herman, 1993, Sliwa, 2016, Johnson King, 2020] claim that worthy action is motivated by rightness itself.

Many find RMF views intuitive, claiming that there is something cold and strange, perhaps even fetishistic, about a person who is motivated by *rightness* instead of by concrete morally important things – like their promise, or the welfare of others [e.g. Smith, 1994]. But RMF views face a major problem: An agent could be motivated by right-making features, yet only *accidentally* or *coincidentally* do the right thing. Such coincidentally right action isn't worthy. This concern is the main motivation for RI views. Markovits [2010, p. 206], for example, notes that (Kant's version of) the RI 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.'

Consider, for example, this (very slightly adapted) case from Sliwa [2016, p. 398].

Bad Jean Jean's friend missed her bus to work and frets over being late to an important meeting; coming late would be a great embarrassment to her. Wanting to spare her friend a major embarrassment, Jean gives her a ride. Let's assume that giving her friend the ride is the right thing to do in these circumstances and the fact that it spares her friend a major embarrassment makes it right. However, Jean would murder her friend's ex-boyfriend if that was the only way to spare her friend embarrassment.[3]

As we will see, we can fill out this case in a few different ways. But in at least some of these ways it seems that Jean does not act with moral worth. She is, in the relevant ways, disconnected from the moral facts. She's just lucky that her laser-focus on saving her friend embarrassment happened to line up with what was right. So, Jean is motivated by what makes her action right but doesn't act with worth – seemingly contrary to the RMF approach.

In this paper I defend a novel version of the RMF account that avoids such accidentality and coincidentality worries.

I motivate my account in two ways. Firstly, the concepts of fluke, accident and coincidence that are central to the moral worth literature play important roles across philosophy. So I consider how these concepts work in other domains. The view I defend follows from an attractive general view about the nature of coincidence.

Secondly, we can add conditions to RMF views in order to rule out coincidentally right actions but many ways of doing this threaten to make the account of worthy action too demanding. My account, I argue, finds it's way between these pitfalls of coincidentality and demandingness.

In section 1 I'll discuss the relation between accident and coincidence. In section 2 I'll consider various approaches to coincidence – arguing that explanatory approaches are preferable. In section 3 I'll discuss how to formulate the RI view in light of this. In sections 4-6 I'll formulate and motivate a novel version of the RMF view. In sections 7 and 8 I'll discuss how this RMF view deals with commonly discussed cases and how it avoids coincidence without making the conditions on moral worth too demanding. Section 9 concludes.

[3]This last sentence differs from Sliwa's version of the case – she does not assume Jean would kill the ex-boyfriend but merely notes that possibility.

1 Accidents and Coincidences

The concept *accident* is part of a cluster of concepts that aren't typically distinguished in the literature on moral worth. It will help to focus on one of these concepts: *coincidence*.

Many kinds of things are called accidents. The concept of coincidence is more tightly circumscribed. The literature suggests that there are two parts to a coincidence. Firstly, there are two or more component events that match in a *striking* way.[4] Secondly, those matching component events are, in some sense, not properly connected (e.g. Hart and Honoré [1985, p. 74], Lando [2017], Bhogal [2020], Berry [2020]).

Some examples:
I toss a fair coin twenty times and it lands heads every time. There is a striking match between the component coin tosses but they seem unconnected. So, it's a coincidence.
You end up on a cruise with your old enemy [Owens, 1992], so there is a striking match between component events – your location and your old enemy's location. If those events are not properly connected then it's a coincidence.
The 'coincidence problem' in cosmology is that there is a striking match between the amount of energy in the universe coming from dark matter and the amount that is dark energy, but these quantities don't seem connected [Bhogal, 2020, p.677-8].
Moral debunking arguments often claim that there is a striking match between our moral beliefs and the moral truth – we have true moral beliefs – but given robust moral realism this is a coincidence since our beliefs and the truth would not be properly connected [Field, 1996, Street, 2006, among many others].

Questions about moral worth fit this structure. Jean gives her friend a ride to work. That was the right thing to do. There is a striking match between the action performed and what is right. But are those facts connected in the appropriate way? If they are not, and Jean only coincidentally did the right thing, then her action is not worthy. So I'll largely focus on the issue of how to rule out coincidentally right actions.

2 Approaches to Coincidence

Again, a coincidence is a striking match between component events that are not properly related. But what is it to be 'properly' related? What relation dispels coincidence? Let's consider two *prima facie* plausible approaches

2.1 The Modal Approach

The first approach is modal – a striking match between events A and B is non-coincidental if certain modal conditions hold. Take the twenty coin tosses. It could *very* easily have been the case that not all landed heads. Had spun the coin a fraction more on the fourteenth toss, for example, it would have landed tails.

[4] See Baras [2022] for a comprehensive recent discussion of strikingness. Luckily, giving an account of strikingness won't be necessary here.

So, perhaps, a matching between events is non-coincidental if it has a kind of modal robustness – it could not easily have failed to hold. This condition is closely related to safety conditions designed to rule out luckily true beliefs. Roughly, my belief in a proposition is safe if I could not easily have falsely believed that proposition.

However, the modal view faces the problem of *modally robust coincidences*. Consider mathematical coincidences. It's a coincidence that the numbers 31, 331, 3331, 33331, 333331, 3333331 and 33333331 are each prime – 333333331 is not prime [Lange, 2010]. But this coincidence could not easily have failed to hold – it is necessary.

Similarly, we can imagine coincidences that hold with the necessity of the laws of nature. Consider this case from Bhogal [2023]:

> **Protons and Electrons** Protons are positively charged. Electrons are negatively charged. However, the absolute value of their charge is the same. Specifically, protons have a charge of $1.602176634 \times 10^{-19}$ coulombs, while electrons have a charge of $-1.602176634 \times 10^{-19}$ coulombs.

If it were a basic law of nature that protons have that charge and, separately, a basic law of nature that electrons have that charge then the matching would be nomically necessary. But still, absent a further story, the matching would be a coincidence.

So, modal robustness isn't enough to dispel coincidence. Maybe the solution is to add another modal condition. We have been considering a condition closely related to *safety* conditions on knowledge or justification. This suggests adding a condition in the spirit of *sensitivity*. Roughly, my belief in a proposition is sensitive if had the proposition been false then I would not have believed it [Nozick, 1981].

But sensitivity-style conditions don't help with the cases just considered. In fact, they are hard to even make sense of. If 31 hadn't been prime would 331, 3331 and 33331 not be prime? There seems to be nothing substantive here. It doesn't seem reasonable to rest the distinction between coincidence and non-coincidence on such questions.

There is much more to say about safety and sensitivity-like conditions in necessary domains. But this would get us too far off track. It is more useful to, now we have noted that robust coincidences make the modal approach to coincidence look uncompelling, note that modal criteria seem particularly unsuitable for developing an account of moral worth.

2.1.1 The Pertinence Constraint

Some views of moral worth are modal, in the sense that the moral worth of an action depends upon what the agent would have done in other circumstances. But, as is often noted, such views seem to inappropriately conflate the moral worth of an action with a broader judgment of the agent's character. The action of a fanatical dog-lover who risked her life to save strangers is still worthy even if, were her dog in danger, she would have saved the dog instead [Markovits, 2010, p. 210]. That she would have saved the dog bears upon her character but does not warrant withholding credit for her heroic act. Similarly, Sliwa [2016, p. 399-400] claims that 'when deciding whether to give an agent credit for an action…we are interested in the motivations that in fact led the agent

to act.' Isserow [2019] calls this idea the *pertinence constraint* – only the motives that *actually* led to action are relevant for moral worth, not counterfactual ones.

Importantly, counterfactuals can be *evidentially* relevant since how an agent would act in other cases can reveal their actual motives. But the pertinence constraint tells us that moral worth isn't directly *determined* by such counterfactuals. Consequently, it rules out using modal approaches to coincidence as part of an account of moral worth.

The pertinence constraint arises from sharply separating the moral worth of an action from a broader evaluation of the agent's character. But there is a more far-reaching and ambitious way to view it. We can see it as part of a broader 'postmodal' approach to philosophy – an approach that has been influential in recent metaphysics. One central thread of the postmodal approach is that the key metaphysical questions are not about modal facts, rather, they are questions about the structure of the actual world. Modal facts are 'are often epiphenomenal, a mere reflection of deeper postmodal structure' [Sider, 2020, p. 3]. So physicalism, for example, shouldn't be understood as a claim about supervenience – this modal fact is a mere symptom of actual world facts, perhaps about what grounds what or about which metaphysical laws hold (e.g. Kim [1993, p. 167], Schaffer [2009, p. 364]).

It's natural to extend this approach to normative domains. What matters for what you should do; what you should believe; and so on, is the structure of the actual world. Facts about other possible situations are not directly relevant. Determining whether this postmodal thought can be developed in a thoroughgoing way is a vast project, but the pertinence constraint can be seen as part of it – as a commitment to grounding moral evaluation in the actual.

* * *

Cases of robust coincidences give us reason to reject the modal approach to coincidence and the pertinence constraint gives us reason to think that it shouldn't, in any case, be applied in accounts of moral worth.

2.2 The Explanatory Approach

A more attractive approach takes coincidence to be an explanatory notion. The rough idea is that in a coincidence the component events are explanatorily 'disconnected' [Owens, 1992, Lando, 2017, Bhogal, 2020]. The twenty coin tosses that all landed heads seem explanatorily independent – hence the sequence was a coincidence.

This approach has no problem with modally robust coincidences – there can be a modally robust matching between events that are nevertheless explanatorily disconnected. Consider **Protons and Electrons**. The theory that posits separate basic laws governing the charge of the proton and the electron implies both that the charges are explanatorily disconnected and that the matching between them is nomically necessary.[5]

[5]Applying this account to mathematical coincidences requires a story about mathematical explanation. See Mancosu [2018, sections 4-7] for a survey.

Further, an explanatory approach to moral worth doesn't violate the pertinence constraint. The actual explanatory connection between action and rightness matters – other possible situations aren't directly relevant.[6]

Giving an account of 'explanatory disconnection' is a difficult task though. We will come back to that in section 4. But a broadly explanatory approach to coincidence is plausible.

* * *

One, perhaps obvious, point before we move on. The coincidence-dispelling relation can't be merely correlational. That is, we can't show some matching between events to be non-coincidental by just pointing to other matchings or correlations. We can't show a matching between fact A and B to be non-coincidental by pointing to a matching between A-type facts and B-type facts more generally, or by pointing to a matching between A and C. The problem, of course, is that these latter correlations could themselves be coincidental. Correlations or matchings cannot, on their own, dispel coincidence.

3 Formulating the RI account

This discussion of coincidence has implications for moral worth. Our focus is on the RMF account, but it will help to briefly see how things play out in the simpler context of the RI view.

The RI account's slogan is that morally worthy action is motivated by the action's rightness. However, the phrase 'motivated by the action's rightness' is ambiguous. On one reading an agent being motivated by ϕ's rightness requires some connection between the agent's action and the actual fact of ϕ being right. But this is not the most common reading in the moral worth literature. The alternate reading is that an agent is motivated by ϕ's rightness when it's part of the content of the agent's motivational state to ϕ that ϕ is right. (One piece of evidence for this claim about the literature comes from commonly discussed cases where an agent is trying to do the right thing but makes some moral mistake. However, they luckily do the right thing, so their action is not explained by the actual rightness of the action. Such cases are often thought of as problems for the slogan that an action has moral worth if it's motivated by the actions rightness – but this only makes sense on the latter, weaker, reading of what motivation consists in.[7])

In the bulk of the paper I'll understand motivation in this latter, weaker, way since it fits better with the moral worth literature and it will allow us to more clearly see the differences between modal and explanatory views. But this is purely stipulative, I'm not meaning to commit to any view in the debate about what motivation really consists in.

[6] A complication: There are accounts of explanation where explanations are determined (in part) by modal facts. The spirit of the pertinence constraint is that the actual situation is what's relevant for moral worth. Does that rule out modal facts playing *any* role in grounding the relevant features of the actual situation? This is a difficult question, one that I'm not able to take on here. But, the pertinence constraint suggests that we should reject accounts of moral worth which appeal *directly* to modal considerations and not via those modal considerations determining explanatory facts.

[7] See, for example, Johnson King's [2020] 'Promise Keeping' case; Singh's [2020] 'Moving' case; the 'Eichmann' case discussed by Arpaly [2002] and Sliwa [2016].

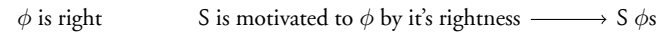

Figure 1: The correlational RI view

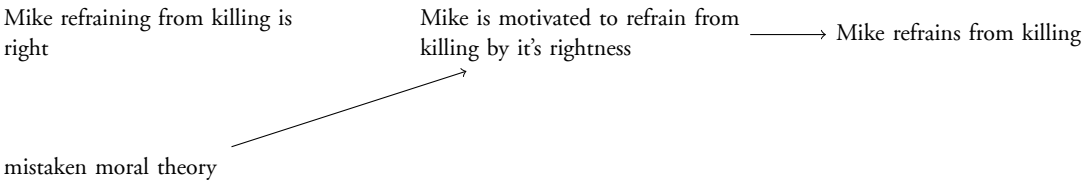

Figure 2: Valuing Suffering

Understood in this weaker way, the slogan that morally worthy action is motivated by the action's rightness doesn't postulate any modal or explanatory connection between an agent performing an action and the action's rightness. Rather, it merely postulates a matching between the agent's motivational state and the moral facts. Figure 1 illustrates the structure.

In the figure the arrows represent explanatory connections. There is an explanatory connection between S's motivation and their action but no connection postulated between the actual fact of ϕ's rightness and the motivation. So, there is a matching between action and rightness, but no connection is postulated between them.

This view is subject to counterexamples:

Valuing Suffering Mike refrains from killing someone for fun because he thinks that refraining is right. However, he only thinks that because he holds very strongly a deeply mistaken moral theory where he values suffering and thinks that people should be kept alive longer so they have more opportunities to suffer.[8]

Figure 2 illustrates this structure. (Remember that we are understanding 'motivation' in the weaker way described above.)

Mike holds an an incorrect moral theory that luckily yields the right result in this case. His action is so disconnected from the actual moral facts that it's a coincidence that he did the right thing.

This case is also a problem for the modal RI view. Consider a natural formulation of the modal RI view: S's ϕ-ing has moral worth if and only if S is motivated by ϕ's rightness and S could not easily have acted wrongly with respect to ϕ. Mike does the right thing coincidentally even though he could not easily have acted wrongly, since it could not easily have been the case that he killed people for fun – because he holds his moral theory so strongly – nor that killing people for fun was acceptable.

Adding a sensitivity-style condition doesn't help. One might say that Mike's action lacks worth because he would still refrain from killing even if the moral facts were different and it was ok to kill for fun. But this doesn't distinguish Mike's action from worthy actions. Imagine someone who refrains from killing someone for fun because they think that refraining is right. They think that because they hold very strongly a moral theory that

[8]This case in the spirit of other 'mistaken theory' cases mentioned in footnote 7.

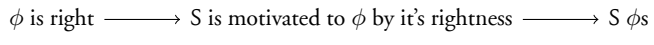

Figure 3: The Explanatory RI view

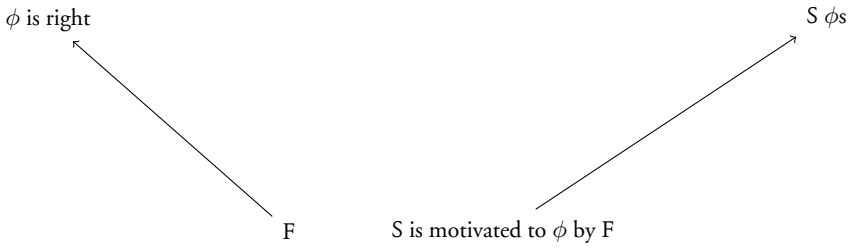

Figure 4: The Correlational RMF view

values human life. Plausibly, this person would not act differently even if the moral facts were different and it was ok to kill for fun. But that doesn't mean their action lacks worth.

The modal RI approach faces such counterexamples because it builds in a modal account of coincidence.

An explanatory version of the RI view (figure 3) looks more promising. If there is a direct explanatory connection between the rightness of an action and the agent performing it then this plausibly makes it non-coincidental that the agent did the right thing. There are further interesting questions about the details of an explanatory RI view but let's move to our main focus – the RMF view.

4 Coincidence and the RMF account

I argued that the RI account should be developed in an explanatory way since the modal approach to coincidence is flawed. The RMF account should also be developed in an explanatory way, for very similar reasons. In this section and the next I'll formulate a novel RMF account, based on explanatory considerations and designed to avoid coincidence worries.

But it's best to start by seeing the problems that simpler RMF views face. The RMF slogan is that morally worthy actions are motivated by the features of the action that make it right.[9] Just as with the RI view, one understanding of this does not postulate any connection between the rightness of the action and it being performed. Take some right action ϕ. What I'll call the *correlational RMF view* says that the agent ϕ-ing has moral worth if (i) the action has some feature, F, that makes it right and (ii) the content of their motivation to ϕ is that ϕ has feature F. This structure is illustrated in figure 4. Again, the arrows represent explanatory connections.

Here, the action's rightness is explanatorily disconnected from its being performed. This makes the correlational RMF view subject to counterexamples. For example:

[9] Of course, there is normally not just one right-making feature of your action [Fogal and Worsnip, 2024]. So there is a hard question about how many, and which, of the right-making features you must be motivated by. This question is largely orthogonal to the issues about coincidences we are considering though it will come up again in section 8.

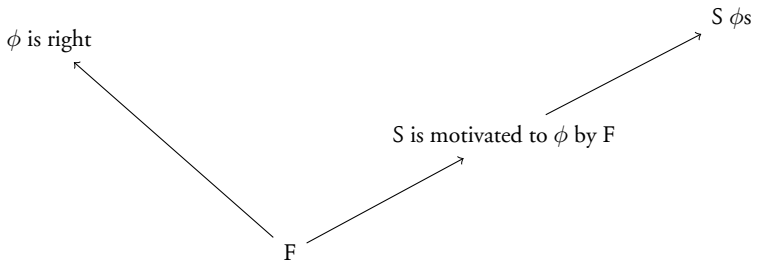

Figure 5: The Third-Factor RMF view

Hiring Stephanie is a hiring manager and gives Yi-joon a job. The content of Stephanie's motivation is that Yi-joon is the most skilled programmer. Yi-joon is the most skilled programmer and that is a good reason to give them the job. However, Stephanie only believes that Yi-joon is most skilled because of incorrect racial stereotypes.

Stephanie's action doesn't have moral worth since her motivation to give the job to Yi-joon because he is the most skilled programmer is disconnected from the actual fact of him being the most skilled programmer.

The obvious fix is to require an explanatory connection between the actual right-making feature and the agent's motivation. This structure is shown in Figure 5. Call this the *Third-Factor RMF view*. This view posits an indirect explanatory connection between ϕ's rightness and the agent ϕ-ing since there is a common explainer of those facts.

The influential views of Arpaly [2002] and Markovits [2010] are naturally interpreted either as Correlational RMF views, or as Third-Factor views. Markovits [2010, p. 205] for example, says that 'my action is morally worthy if and only if my motivating reasons for acting coincide with the reasons morally justifying the action'. Arpaly's [2002, p. 84] account of whether is action is worthy 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' (though she allows modal considerations to be relevant for the *degree* of moral worth). There's an interesting interpretative question here but both views are probably most charitably understood as third-factor views.

4.1 Jean

Unfortunately, even though the Third-Factor view postulates an indirect explanatory connection between rightness and the agent's action it still faces coincidentality problems. Consider Bad Jean again. Jean gives her friend a ride motivated by saving her embarrassment. Saving her friend embarrassment makes it right to give her friend a ride. But (in at least some versions of the case)[10] Jean does not act worthily. This looks like a counterexample to the Third-Factor RMF view.

It might seem easy to deal with this. One might suggest that Bad Jean doesn't have moral worth because she would act terribly in closely related circumstances – she would kill the ex-boyfriend. But, as Sliwa [2016, p.

[10] I'll come back to this qualification soon.

399-400] notes, such counterfactuals can't determine moral worth – that would violate the pertinence constraint and confuse the moral worth of this action with a judgment of Jean's character. And, I would add, in the spirit of section 2.1, that it would build in an inadequate modal conception of coincidence.

But, you might protest, wasn't the whole point of the Jean case about counterfactuals? Doesn't Sliwa think that the reason Jean doesn't act worthily is because her right action is so 'precarious' – it could so easily have gone wrong? So wasn't Sliwa's, and our, reasoning that Bad Jean doesn't act with worth a violation of the pertinence constraint? I don't think so. The charitable way to understand Sliwa's appeal to counterfactuals is as merely evidential. What Jean would, or might, do in other cases can be evidentially relevant to, but cannot determine, moral worth. Anyone, like Sliwa and me, who accepts the pertinence constraint, must think that Jean's counterfactual killing of the ex-boyfriend is merely evidentially relevant to whether giving her friend a ride has moral worth.[11]

So, **Bad Jean** continues to be a counterexample to the Third-Factor RMF view. However, notice that some variants of Jean do intuitively act with worth. If Jean is motivated to give her friend a ride to save her embarrassment but wouldn't murder the ex-boyfriend, and would generally act in a reasonable way in related situations, then RMF theorists are inclined to judge her action worthy.

One way to understand Sliwa [2016, section 3] is that she leverages this into an argument against RMF views generally. Jean is motivated by saving her friend embarrassment. In some versions of the case she intutively doesn't act worthily; in others, she intuitively acts worthily. But the only difference between these cases is in what Jean would do in other circumstances, and this is not relevant for determining moral worth. So, Sliwa concludes, Jean doesn't act worthily in the seemingly 'good' cases either. So, RMF views are just on the wrong track – being motivated by right-making features is not important for moral worth. (This interpretation goes a little beyond Sliwa's text – whether this is precisely Sliwa's intended argument is not totally clear to me.)

This is a powerful challenge. In response the RMF theorist must distinguish between variants of the Jean case, distinguishing worthy from non-worthy actions, without violating the pertinence constraint. In line with our general approach to coincidence we must draw the distinction in *explanatory* rather than modal terms.

In the next section I'll do that, formulating a novel RMF view.

5 The anti-coincidence condition

In order to formulate an account of moral worth that avoids coincidentality we need to answer a general question: When we have a striking match between facts A and B what explanatory relation between them dispels coincidence?

Notice that it's not enough to separately explain A and B, and then 'staple together' these two explanations. Consider again the 20 coin tosses that landed heads. We could, at least in theory, give a microphysical explanation of the first toss landing heads – appealing to the precise spin and velocity it was thrown at, and so on. We could

[11] This is why I have been allowing that Bad Jean might act worthily in some versions of the case. As I'll argue later, Jean's counterfactual killing is good, *but not conclusive*, evidence that she acts unworthily in the actual case.

also do this for all the other coin tosses and then conjoin those twenty explanations. But merely stapling together these distinct explanations doesn't dispel coincidence.

This 'anti-stapling' thought is visible in a variety of debates. For example, in the debunking literature some point out that, given moral realism, we can separately explain what's morally right and why we have the moral beliefs that we do but it can still seem coincidental that we have true moral beliefs. Street [2016, p. 31] expresses this idea:[12]

> One may explain each side of the coincidence in as much depth as one likes – going into wonderful normative depth about why family and friendship are valuable, and wonderful scientific depth about why we were selected to think this. But all this goes nowhere toward explaining the thing that really needs to be explained, namely the coincidence itself.

This anti-stapling thought is why the **Correlational RMF** view fails, since it merely conjoins separate explanations of S ϕ-ing and ϕ being right.

Notice though, perhaps surprisingly, that identifying a common explainer of A and B isn't enough to dispel coincidence either. For example, given certain reasonable background assumptions, the Big Bang may be a common explainer of any two physical events [Owens, 1992, p. 8]. But that doesn't mean that there are no coincidences. Similarly, Bhogal [2022] discusses a slight variant of a case from Lando [2017] where a child throws two balls up in the air, they collide and fly off in different directions and both hit the high-A on different pianos, that happen to be close by. It's clearly a coincidence that the same note was hit, even though the collision of the balls is a common explainer.

Again, this point plays an important role in the debunking literature. Some respond to debunking arguments by arguing that it's not a coincidence that our moral beliefs are true (even given moral realism) since we can identify common explainers of our moral beliefs and the moral truth [e.g. Enoch, 2011, Skarsaune, 2011, Wielenberg, 2010].

Consider an example of this strategy from Korman and Locke [2020]. I believe that I should feed my child. Why? Because it helps them to survive, and natural selection instills in me beliefs that help my children survive. It's morally good to feed my child. Why? Because it helps them to survive, and survival is morally valuable. So, that feeding helps children survive is a common explainer of my belief and the moral fact.

But, many people (e.g. Korman and Locke [2020], Faraci [2019], Lutz [2018], Noonan [2023]) think that identifying this common explainer doesn't help. It still seems that the factors that led to my moral beliefs *just so happened* to line up with the factors that explain the moral truth.

On it's own, identifying a common explainer of A and B isn't enough to dispel coincidence. This is why the **Third-Factor RMF** view fails. Even without looking at cases like Bad Jean we could have rejected the third-factor view on these very general grounds – a common explainer of S ϕ-ing and ϕ being right isn't enough to dispel coincidence.

[12] See also Field [1996, section 5], Linnebo [2006, section 2], Berry [2020, section 5b].

So what is enough to dispel coincidence? A direct explanatory connection between A and B – A explaining B or B explaining A – would be enough, but it isn't necessary. I dispel the coincidence of twenty coin tosses landing heads by telling you that the coin was tossed by an extremely precise coin-tossing robot that I programmed to toss heads. But this is to identify a common explainer of the tosses landing heads, not to say that the first coin landing heads explained why the second landed, and so on.

A better story about the coincidence-dispelling relation involves developing the 'anti-stapling' intuition. A way of putting the intuition is that we don't just want an explanation of A *together with* an explanation of B; we want to explain the *matching* between A and B. Following Lando [2017] and Bhogal [2020] we can distinguish between two facts in cases like the twenty coin tosses. There is the fact that all of the tosses landed *the same way* – Bhogal calls that the *matching proposition*. And there is the fact that coin toss one landed heads, coin toss two landed heads…and coin toss twenty landed heads – Bhogal calls this the *particular proposition*. Similarly, in the Piano case mentioned above, we can distinguish the fact that the *same note* was struck on both pianos – the matching proposition – from the fact that the high-A was struck on the first Piano and the high-A was struck on the second piano – the particular proposition.

Given this setup, one way to put the anti-stapling intuition is that we don't want to just explain the particular proposition, we want a distinct explanation of the matching proposition. We don't want to just explain why coin toss one landed heads, coin toss two landed heads…and coin toss twenty landed heads; we want a distinct explanation of why all the tosses landed the same way. For example, that I was trying to test the accuracy of my coin-tossing robot can explains why all the tosses landed the same way without merely explaining why each toss landed heads separately.

This points us towards an anti-coincidence condition. When there is a striking match between A and B we want an explanation of the distinctive way in which A and B match – one that isn't merely an explanation of A and B together. We want, that is, an explanation of the matching proposition that isn't merely an explanation of the particular proposition – call this a *unified* explanation.

The way we originally put the anti-stapling intuition is that merely stapling together separate explanations doesn't dispel coincidence. Now we can clearly see what this means. Unified explanations dispel coincidence. Explanations that are not unified, in this sense, are the ones that seem 'stapled together'.

(As the Pianos case reveals to us, an explanation of A and B can fail to be unified even when there are common explainers of A and B. When we explain why the two balls hit the same note on both Pianos the best we can do is explain the specifics about why the first ball hit the high-A on the first piano and the second ball hit the high-A on the second piano. There is no distinctive explanation of the *matching*. And this is true even though the collision of the balls is a common explainer (see Bhogal [2020, section 5]).

Unified explanations are what we need to dispel coincidence. I take this anti-coincidence condition to be in the spirit of Lando [2017], Bhogal [2020] and Heering [fort.].

I'll say more about the concept of unified explanation in the next section. But we can now see what an RMF account has to look like to avoid coincidentality worries.

Unified Explanation RMF: For an agent's doing ϕ to have moral worth (i) the agent must be non-instrumentally

motivated by right-making features of ϕ[13] and (ii) there must be a unified explanation of why the agent did the right thing – that is, an explanation of *why the agent did the right thing* that is not just an explanation of *why the agent did ϕ and why ϕ was right*.

Condition (i) is what makes it an RMF view; condition (ii) is designed to rule out the agent coincidentally doing the right thing.[14]

In this section we saw that general considerations about avoiding coincidence motivate the **Unified Explanation RMF**. Soon, I'll argue that the view gets attractive results about disputed cases in the moral worth literature. Just before that, it will be helpful to say more about unified explanations in the moral case.

6 Some Unified Explanations

The view I've suggested involves the somewhat unfamiliar concept of a *unified explanation*. This is cashed out in terms of the more general concept of *explanation*. *Explanation* is familiar but we are, perhaps, unaccustomed to it playing a key role in this debate. It's more common to see accounts of moral worth given in terms of concepts like *knowledge* or based on claims about *counterfactuals*. Explanation is a disputed notion but, of course, so are knowledge and counterfactuals. So it's fair for me to give my account of moral worth in terms of explanation. But still, it will be helpful to give some examples of unified explanations of an agent doing the right thing.

This will be helpful for another reason too. If an agent does the right thing non-coincidentally there is a *unified* explanation of why the agent *did the right thing*, not just an explanation of why the agent did ϕ and why ϕ was right. But how is this possible without collapsing into an RI view where the agent is motivated by rightness itself?

So I'll mention a few models for what a unified explanation can look like – there are more that I won't mention. I think it's likely that different models will hold in different cases. In one case the first model might hold, in another perhaps the third model holds. This is fine and both actions can be worthy. The point, again, is just to help the reader get a grip on what kinds of things count as unified explanations.

(1) Perhaps the agent is motivated directly by right-making features but has a kind of background attentiveness to rightness itself which acts as a 'filter' 'plac[ing] limits upon an agent's capacity…to act on other motives' [Isserow, 2021, p. 281] so that motives that would lead to immoral actions are filtered out. This would, for example, prevent Jean from acting out of the motive of saving her friend embarrassment when that motive would lead to murdering the ex-boyfriend. This filter would generate a relevant explanatory connection between rightness and the agent's action. (Stratton-Lake [2000] is plausibly understood as requiring that such filter is present for an act to have worth.)

[13]The non-instrumentality condition is added to rule out cases like the politician who is motivated to help people in need, but only cares about that because being seen to help people in need will benefit his electoral chances. (see e.g. Markovits [2010, p.230]. Isserow [2019, section 2])

[14] In footnote 1 I mentioned a version of my view that assumes that worthy actions are subjectively right, not objectively right. It's easy to see how this view would go – instead of requiring a unified explanation of the matching between and agent ϕ-ing and ϕ being objectively right we require a unified explanation of the matching between and agent ϕ-ing and ϕ being subjectively right. But evaluating the plausibility of such a view is a task for another time.

(2) Perhaps the agent is motivated directly by the right-making features of an action, but the explanation for why the agent is motivated by those features is that they have some direct insight into the moral facts. Such an agent might just form a desire to save their friend from embarrassment, for example. But, ultimately, that desire is formed because of some insight they have into rightness, even if they don't conceptualize it as such, even if they don't think of their action as morally right, and even if they don't think of their action in moral terms at all. This insight would provide a unified explanation for why the agent did the right thing.

Roughly what I mean by 'moral insight' is a kind of epistemic connection to certain facts, whereby they can explain our desires and actions but without belief about those facts. Outside of the moral realm this type of insight is very common. As I move around the world facts about my center of mass are part of what explain my desires and actions. It's not a coincidence that I typically move in a way that keeps the base of my body underneath my center of mass. I have insight into the facts about my center of mass, even if I don't have any thoughts about them.

There are lots of facts that explain our desires and actions and, intuitively, we have insight into even if such facts don't 'register' with us. You might find someone physically attractive, and this is *because* of facts about the symmetry of their face, even if you don't have any thoughts about the symmetry of their face. If someone asks you later whether their face was especially symmetric you wouldn't have anything to say. But you do seem to have access and insight into facts about their facial symmetry and this (partially) explains your desires and actions. A huge amount of our day-to-day actions seem to have a similar structure.

Distinctively moral insight is more complicated since it's unclear what moral facts are. Assuming that there are moral facts they are either natural or non-natural. If they are natural then they will be causally efficacious and so it's completely possible that we can have insight into those facts and they can explain our desires and actions in roughly the way described above. If the moral facts are non-natural and not causally efficacious then it's a little harder to see what moral insight can come to, but only because it's correspondingly harder to see how we could have any kind of access to those facts in the first place. Of course, now is not the time to discuss such access worries. But hopefully the reader can see how moral insight of this form could generate a unified explanation of why the agent did the right thing.

(3) Similarly to the last option, perhaps the agent is motivated directly by the right-making features of an action, but the reason that the agent is motivated by those features is because they have gone through some reliable process of moral education. The agent may have been brought up to be kind and generous and save their friends from embarrassment where possible. Since their moral education is reliable the actions of this agent will be connected to facts about rightness. Presumably, the reason that the agent had a reliable moral education involves, at some point in the causal chain, a connection to rightness – perhaps on the part of their teachers, or their teachers' teachers. So, there is a relevant connection between rightness and the agent's action.

In fact, this is a very common case. An agent acts generously because of their moral character, and this moral character is due to their moral upbringing. But an agent need not be motivated by rightness itself, or even think of their action in moral terms, when they are acting.

(1), (2) and (3) illustrate ways in which there could be a unified explanation of the agent doing the right thing without the agent being motivated by rightness itself. Notice that we can have a connection between the agent's

action and rightness without the agent knowing, or even believing, that what they are doing is right.

Further, these examples illustrate a feature of my account: It can sometimes be hard to know, even for the agent themselves, whether an action has moral worth. This is because it can be hard to know whether your action is explanatorily connected to the actual moral facts. But that moral worth isn't always easy to judge is not a problem, I take it, for an account of moral worth.

There are clearly other possibilities for this connection between rightness and action – the few models I mentioned here shouldn't be taken to be definitive of the account. But our investigation of coincidence suggests that there needs to be some such connection for an agent to act non-coincidentally rightly.

7 Some Cases

In this section I'll argue that Unified Explanation RMF deals well with some commonly discussed cases – both cases normally taken to favor the RMF view and those taken to favor the RI view. In particular, it elegantly and plausibly distinguishes between variants of the cases.

7.1 Huck Finn

The most discussed case in the moral worth literature is that of Huckleberry Finn – the character from Mark Twain's novel. Huck is a white teenager living in the south of the USA in the mid-19th century. He befriends an escaped slave, Jim. At a key point he is conflicted about whether to turn Jim in or to help him escape. He ends up helping Jim escape even though he thinks it is morally wrong since it amounts to stealing from Jim's 'rightful owner'.

The case is a core motivation for the RMF view. It's clear, many think, that Huck helping Jim escape is worthy even though he is not motivated by the thought that it is right. What matters is that Huck is motivated by the right-making features of his action – Jim's humanity and the value of his life.

But, I've argued, it's not enough that Huck is motivated by right-making features – there needs to be an explanatory connection between the facts about rightness and Huck's action. And in a natural interpretation of the case there is such an explanation. Huck, it seems, doesn't value Jim's humanity at random. It is not, we want to say, merely coincidental that Huck's valuing Jim's humanity lines up with what really is of value. Rather, Huck has some moral insight that explains his motivations and his ultimate choice to help Jim.

There are versions of the case where Huck doesn't have such a connection to the moral facts. Sliwa [2016, section 8], for example, describes a version of the case where Huck doesn't turn Jim in because of their friendship. But, of course, the morally relevant fact isn't that Jim is Huck's friend. In this case Huck doesn't seem connected to the moral facts and it does seem coincidental that Huck did the right thing – the Unified Explanation RMF gets that result.

But some versions of the case involve a connection to the moral facts. Here, for example, is part of Arpaly's [2002] description of the case:

> during the time he spends with Jim, Huckleberry undergoes a perceptual shift…Talking to Jim about his hopes and fears and interacting with him extensively, Huckleberry constantly perceives data (never deliberated upon) that amount to the message that Jim is a person, just like him. Twain makes it very easy for Huckleberry to perceive the similarity between himself and Jim: the two are equally ignorant, share the same language and superstitions, and all in all it does not take the genius of John Stuart Mill to see that there is no particular reason to think of one of them as inferior to the other. While Huckleberry never reflects on these facts, they do prompt him to act toward Jim, more and more, in the same way he would have acted toward any other friend…As mentioned above, Huckleberry is not capable of bringing to consciousness his nonconscious awareness and making an inference along the lines of 'Jim acts in all ways like a human being, therefore there is no reason to treat him as inferior, and thus what all the adults in my life think about blacks is wrong.' (pp. 76-77)

The natural way to read this, at least for me, is that, during his time with Jim, Huck gains access to moral facts. Arpaly suggests that the fact that 'there is no particular reason to think of one of them as inferior to the other' 'prompts' Huck to act in certain ways towards Jim. And, she seems to say, Huck has a 'nonconscious awareness' of something like the fact that 'Jim acts in all ways like a human being, therefore there is no reason to treat him as inferior, and thus what all the adults in my life think about blacks is wrong.'

The picture, it seems, is that Huck gains some loose, inchoate, access to the moral fact that black people aren't inferior and so should be treated equally. And this is (part of) what explains his actions. At least, this is the picture I get from Arpaly's passage that convinces me that Huck's action has moral worth.

There's another reading of the above passage, though, where Huck has no connection to the moral facts but rather gains access to purely descriptive facts. On this reading his 'nonconscious awareness' is of merely descriptive facts about the similarity between him and Jim, or black people more generally, in various descriptive respects. This is, in fact, the reading that fits with Arpaly's official view. Her official view does not require that Huck has any connection to the moral facts in order for his act to be worthy. However, she is at her most convincing when you get the sense that Huck has manged to gain contact with the moral realm.

My account says that on the interpretation where Huck has some kind of unconscious connection to the moral facts, his action is worthy. This, it seems to me, is the right result.

Notice that my view does not require that Huck has knowledge of, or even believes, the moral truths. I agree with Arpaly that it this not required. But the moral truths do need to play a role. While Huck 'never reflects on these facts', they must 'prompt him' to act as he does.

7.2 Jean, again

While the Huck Finn case has been a core motivation for the RMF view, the Jean case (along with related cases like Herman's [1981, p. 364-5] art thief) has been taken to be a problem for the RMF approach.

The challenge, as we discussed in section 4.1, is distinguishing between variants of the case where Jean acts with worth and where she does not without appealing to counterfactual differences. How does the Unified Explanation RMF do this?

The starting point was that Bad Jean does the right thing in giving her friend a ride to work and so saving her friend embarrassment, but would kill her friend's ex-boyfriend if that was the only way to save her friend embarrassment. As we noted in section 4.1, anyone who accepts the pertinence constraint has to say that this counterfactual is merely evidentially relevant to Jean's moral worth – so that is how my account will distinguish variants of the case.

The fact that Bad Jean would kill the ex-boyfriend is evidence that her actual action is not appropriately explanatorily connected to rightness and so, on my account, is evidence that she doesn't act with moral worth.

Here's an analogy: I'm a bad chess player. Imagine I, during a game, play knight to h6 and that's the right move. If I, in related situations, would play completely terrible moves then that's strong evidence that in the actual case my move isn't appropriately explanatorily connected to the fact that it's the right move. It's strong evidence that I was just lucky and so don't deserve credit for playing the right move.

But it's not conclusive evidence. It's possible that I normally play terrible moves, but in this case I actually do calculate things correctly, recognizing that my move will trap the opponents rook. Bad chess players, especially ones who are improving, can sometimes do this, and when they do they deserve credit.

Similarly the fact that Bad Jean would do terrible things in related cases is strong evidence that in the actual case her action is disconnected from the moral truth. So that's how my view diagnoses judgements about the case – we have strong evidence that her action isn't worthy.

But it's not conclusive evidence. If there is a unified explanation of why Jean did the right thing then, typically, there will be reason to expect Jean to act rightly in related situations – for example, not murdering her friend's ex-boyfriend to prevent embarrassment. But this won't always be the case. Perhaps there is something about Jean's character that means that her good moral education doesn't properly transmit to action when it comes to killing ex-boyfriends. So there are some cases, my view suggests, where Jean's act of giving her friend a ride is worthy even if she would murder the ex-boyfriend.

This is the right result – the fact that there are some possible situations where an agent would do terrible things doesn't undermine the worth of their actions in general. There are people who *actually* do terrible things and can still act worthily, even if, of course, their character is flawed.

To summarize, Jean's actions are not worthy when her giving a friend a ride is explanatorily disconnected from the moral facts. That she would do very bad things in related circumstances is good but not conclusive evidence for this explanatory disconnection.

Notice, then, that the distinction between cases like Bad Jean and Markovits's fanatical dog lover is that Jean would act so terribly in such closely related situations that we are inclined to suspect that her actual actions are disconnected from the moral truth, while the fact that the dog-lover would save her dog, if the dog was in danger, isn't such strong evidence. Consider the chess analogy again. If I would play completely terrible moves in related

cases, that's good evidence that my actually correct move is lucky – it's disconnected from the actual fact of the move being right. But we don't get such strong evidence when a player plays the right move but would, in certain related cases, make more understandable mistakes.

7.3 Venom

One final case. Here is one from Keshav Singh [2020].

Venom Jack, a surgeon, is hiking when he sees a stranger get bitten by a venomous snake and faint. He immediately makes an incision near the bite so that the venom will drain out. Making the incision is the right thing to do, and Jack's reason for doing it (that it will allow the venom to drain out) is part of what makes it right. But Jack doesn't have any particular concern for doing the right thing in this case, nor does he conceive of his reason as one that makes his action right. He is simply intrinsically interested in draining venom out of wounds.

Intuitively, Jack's action doesn't have moral worth.

But Singh also discusses a variant, **Venom***, where Jack makes the incision because of his intrinsic desire to save lives, not his intrinsic desire to extract venom. It seems much more intuitive to ascribe Jack's action moral worth in this case. Similarly, if Jack acts out of an intrinsic desire to improve the welfare of other people then his action seems perfectly worthy.

But, Singh's view is that such actions are not worthy since they have the same structure as **Venom**. I think this result is implausible, and my view avoids it. In the original version of the case Jack is intrinsically motivated by extracting venom – this is strong evidence that his actions are not explained by the actual moral facts. That intrinsic motivation is unlikely to be the result of genuine moral insight or reliable moral education, for example. But if Jack is instead intrinsically motivated by improving the welfare of others then it's much more plausible that his motivation, and thus his action, is explanatorily connected to the facts about rightness. My view can appeal to evidential considerations to diagnose the difference between **Venom** and variant cases. It's an advantage of my view that it doesn't treat all cases with the same high-level structure as **Venom** in the same way.

* * *

My approach handles cases that have traditionally been taken to favor the RMF view – like **Huck Finn** – and those that have been taken to tell against it – like **Jean** and **Venom**. The appeal to explanatory and evidential considerations can draw the line between versions of the cases where agents act with worth and where they do not.

8 Coincidence vs Demandingness

I've motivated the Unified Explanation RMF by considering the general nature of coincidence and by arguing that the account gets plausible results in a variety of cases.

I'll end by illustrating a general challenge for RMF views: They have to rule out coincidentally right actions, but moves to do this threaten to make the conditions on moral worth too demanding. We can understand many of the arguments in this paper as showing that my account can meet this challenge.

Here, for example, is one natural response an RMF theorist could make to the Bad Jean case: Jean is not motivated by *enough* of the right-making features – more than just her friend's embarrassment is morally relevant. That's why her action isn't worthy. However, most of our actions are only motivated by a very small set of the right-making factors. If I give money to an unhoused person, motivated by empathy, I'm likely not motivated by all of the complex morally relevant factors about the nature of homelessness in my city; what would best alleviate it; which charitable donations are best; and so on. Certainly, we can imagine a ten year old child acting in this way, without holding any complex set of factors in their mind. Of course, this very quick argument is not conclusive, but we can see the threat of the account becoming too demanding.

(A separate concern with this strategy: Even if we imagine a version of Jean who was actually motivated by all of the right-making factors it could still be coincidental that she did the right thing. Imagine, in the actual case, factors A, B, C and D are all the morally relevant factors, and Jean is motivated by those factors. But then, imagine, that Jean is laser-focused on factors A, B, C and D and so is motivated by those factors in every action, big and small, that she performs in her life, even when those factors are not morally relevant at all. Intuitively, she is just lucky that in this particular case her focus on factors A, B, C and D in all aspects of her life lines up with what is actually morally relevant. Consequently, she coincidentally does the right thing.

Demanding that agents must be motivated by *more* of the right-making factors is not, on it's own, a solution to coincidence problems because it is, in effect, just to demand yet more correlations, which could themselves be coincidental. We can add as much correlation as we want between the factors that actually motivate the agent and the factors that make for the rightness of an action, but as we discussed in section 2, more correlations do not dispel coincidence, we need genuine explanatory connection.)

Here is a related strategy. Maybe the problem with cases like Bad Jean is that Jean is not motivated by the *fundamental* right-making features. But people are very rarely motivated by the morally fundamental so adding such a condition threatens to make the view overly demanding [see e.g. Portmore, section 3]. Further, two people who disagree on the fundamental right-making features of some particular action – a utilitarian and a Kantian, for example – can both act with worth.

There is a huge amount more to say about these two strategies, but we can see that it's difficult to walk the line between coincidence and demandingness.

Some other approaches to avoiding coincidence threaten to make moral worth too demanding in another way – by building too much into the mind of an agent who acts worthily.

For example, Singh [2020] claims that 'A right action has moral worth if and only if the agent performs it on the basis of sufficient moral reasons as such.' To be motivated by sufficient moral reasons as such is to act under the 'guise' of those reasons being sufficient moral reasons. And 'to act for some reason under the guise of a moral reason is to be motivated by that reason in virtue of taking it to contribute to the overall moral status of the action' (p.170-172). So, it's not enough to act for the right reasons, you need to take those reasons to be the right reasons.

But Huck Finn, it seems, is not motivated to help Jim in virtue of taking some reason to be a sufficient moral reason to help him. In fact, Huck thinks he has sufficient reason *not* to help Jim. An RMF theorist should not want to rule out Huck acting with worth since getting the intuitive results in the Huck Finn case is a key part of the motivation for the RMF view in the first place.

Singh suggests that his account is not too demanding since Huck can 'tacitly' take Jim's personhood to 'constitute sufficient moral reason to help him' even though he explicitly believes the opposite. I'm not convinced, but now is not the time to pursue this issue. We can see, though, the potential threat of RMF accounts being too demanding and ruling out the worth of Huck Finn's action.

Relatedly, certain epistemic conditions that RMF theorists postulate can easily make their account too demanding. Isserow [2019, p. 262], for example, claims that agents that are motivated by right-making features 'must plausibly be justified in believing that they ought to act as they do' in order to act worthily.[15]

But again, such a condition seems to rule out Huck Finn. Huck doesn't believe that he ought to act as he does. And we might doubt whether he even has justification for such a belief, regardless of whether he holds it, given the strong testimonial evidence he is getting from society as a whole. Though these questions about moral justification are too far afield to get into further.

There is much more to say about both Isserow's and Singh's views, and about other approaches in this spirit. The point is not to refute those views but to illustrate the concidentality vs demandingness trade-off that we face.

Other approaches can be over-demanding in a different way. Lord [2017] gives an attractive account of moral worth, based on *manifesting know-how* about how to use the reasons you have.[16] For Lord this involves having certain dispositions to act that are sensitive to the actual moral facts. For example, he discusses a version of Kant's shopkeeper case, claiming that the shopkeeper should be disposed 'to give $1.00 back [to the customer] when the fact that $1.00 is the correct change provides a sufficient (moral) reason to give $1.00 back' (p. 458). There are a lot of attractions to this view but the major concern is that the appeal to dispositions violates, at least the spirit of, the pertinence constraint. In particular, it seems that his view rules out moral worth for 'out of character' actions – agents who are generally disposed to act badly but who manage, sometimes, to act rightly, and intuitively for the right reasons – and so ends up being over-demanding in a different way.[17]

Of course, this is not close to an exhaustive classification of the possible RMF views and it's not close to a full discussion of the specific views I mentioned. But we can see the trade-off that RMF views face between avoiding coincidentally right actions and being over-demanding.

My approach, as I've argued in sections 4 and 5, avoids coincidentally right action. Section 6 illustrates how the view doesn't build too much into the agent's psychology, and so doesn't rule out Huck Finn from acting with worth. And since it is an explanatory approach, not a modal approach, it, as we discussed in section 2

[15] Isserow actually holds a disjunctive view where either concern for rightness itself or right-making features can make for moral worth. For our current purposes we can focus on the right-making features disjunct of her view.

[16] Cunningham [2021] has a closely related view.

[17] A possible response Lord could give is that relevant dispositions can be extremely fragile. But, if we strip dispositions of what seems to be their core feature – their connection to modality – then it becomes hard to get a grip on what the appeal to dispositions is supposed to do for us.

doesn't violate the pertinence constraint. I think it is successful at walking the line between coincidence and over-demandingness.

9 Conclusion

Morally worthy actions are non-coincidentally right. To make sense of this platitude we must understand what a coincidence is. Doing so motivates an account of moral worth – built upon the idea of a *unified* explanation – and casts doubts upon accounts of moral worth that don't fit with a broader understanding of coincidence.

One last point: Notice that the Unified Explanation RMF is formulated as a *necessary* condition for moral worth. This is because of deviant explanatory chain cases like this: Imagine an agent recognizes that some particular action has right-making features and becomes motivated to do the complete opposite because they are an anti-moralist. Then a demon interferes with their brain and flips their motivation so now they are motivated to perform the action. The agent doesn't act with moral worth even though they are motivated by right-making features and there is a unified explanation (via the actions of the demon) of why they acted rightly.

It's hard to know how worried to be about cases like this, especially since such closely analogous cases can be given for RI views. Imagine an agent recognizes the rightness of an action and then becomes motivated to do the complete opposite because they are an anti-moralist. Then a demon interferes with their brain and flips their motivation so now they are motivated to perform the action. If our view of moral worth is based upon the motivations that an agent has, as both the RI and RMF views are, then it's not surprising that a demon fiddling directly with people's motivations can cause problems.

One reaction is to add some concept to our account that builds in an anti-deviance condition. Many authors already do this. For example, the RI view of Johnson King [2020] involves the idea that an agent must *deliberately* do the right thing. Part of the role of deliberateness is to rule out deviant chains. Similarly with Isserow's [2019] *competently bringing about* and the concept of *manifesting know-how* that Lord [2017] and Cunningham [2021] stress. But such accounts don't really *solve* the deviant chain problem, rather the existence of a solution is built into the accounts.[18] To be clear, this is not a criticism! It's a totally reasonable approach, given that the problem of deviant chains is a deep issue across a variety of debates in philosophy.

I could, similarly, adapt the Unified Explanation RMF by adding such an anti-deviance condition thus stating necessary and sufficient conditions for moral worth. Or I could leave the account as it is. Either way, my account identifies the core of moral worth – it is motivation by the right-making features and the existence of a unified explanation that makes it non-coincidental that the agent acted rightly. That's not merely a necessary condition on moral worth – it's an identification of what matters for moral worth.

The attractions of the Unified Explanation RMF – particularly that it avoids coincidentality worries – is a powerful argument against RI views. It makes it hard, at least for me, to see why we would want to accept an RI view in its place.

[18] For example, Lord [2018] characterizes his discussion of deviant chains by saying 'What I've done is isolated where the problem lies' (p. 138).

References

Nomy Arpaly. *Unprincipled Virtue: An Inquiry Into Moral Agency*. Oxford University Press, November 2002.

Dan Baras. *Calling for Explanation*. Oxford University Press, July 2022.

Sharon E Berry. Coincidence avoidance and formulating the access problem. *Canadian journal of philosophy*, 50 (6):687–701, August 2020.

Harjit Bhogal. Coincidences and the grain of explanation. *Philosophy and phenomenological research*, 100(3): 677–694, 2020.

Harjit Bhogal. What's the coincidence in debunking? *Philosophy and phenomenological research*, July 2022.

Harjit Bhogal. Explanationism versus modalism in debunking (and theory choice). *Mind*, 132(528):1005–1027, 2023.

Joe Cunningham. Moral worth and knowing how to respond to reasons. *Philosophy and phenomenological research*, 105(2):385–405, 2021.

David Enoch. *Taking Morality Seriously: A Defense of Robust Realism*. OUP Oxford, July 2011.

David Faraci. Groundwork for an explanationist account of epistemic coincidence. *Philosophers Imprint*, 2019.

Hartry Field. The a prioricity of logic. In *Proceedings of the Aristotelian Society*, volume 96, pages 359–379, 1996.

Daniel Fogal and Alex Worsnip. What the cluster view can do for you. In Russ Shafer-Landau, editor, *Oxford Studies in Metaethics, Volume 19*. OUP, 2024.

Herbert Hart and Tony Honoré. *Causation in the Law*. OUP Oxford, May 1985.

David Heering. Explanationism about freedom and orthonomy. *The journal of philosophy*.

Barbara Herman. On the value of acting from the motive of duty. *The Philosophical review*, 90(3):359, July 1981.

Barbara Herman. *The Practice of Moral Judgment*. Harvard University Press, 1993.

Nathan Robert Howard. One desire too many. *Philosophy and phenomenological research*, 102(2):302–317, 2021.

Jessica Isserow. Moral worth and doing the right thing by accident. *Australasian journal of philosophy*, 97(2): 251–264, April 2019.

Jessica Isserow. Doubts about duty as a secondary motive. *Philosophy and phenomenological research*, 105(2): 276–298, 2021.

Zoe Johnson King. Accidentally doing the right thing. *Philosophy and phenomenological research*, 100(1):186–206, 2020.

Jaegwon Kim. *Supervenience and Mind*. Supervenience and Mind. Cambridge University Press, Cambridge, January 1993.

Daniel Z Korman and Dustin Locke. Against minimalist responses to moral debunking arguments. *Oxford Studies in Metaethics*, 15, 2020.

Tamar Lando. Coincidence and common cause. *Noûs*, 51(1):132–151, March 2017.

Marc Lange. What are mathematical coincidences (and why does it matter)? *Mind*, 119(474):307–340, July 2010.

Øystein Linnebo. Epistemological challenges to mathematical platonism. *Philosophical studies*, 129(3):545–574, June 2006.

Errol Lord. On the intellectual conditions for responsibility: Acting for the right reasons, conceptualization, and credit. *Philosophy and phenomenological research*, 95(2):436–464, September 2017.

Errol Lord. *The Importance of Being Rational*. Oxford University Press, 2018.

Matt Lutz. What makes evolution a defeater? *Erkenntnis*, 83(6):1105–1126, 2018.

Paolo Mancosu. *Explanation in Mathematics*. Metaphysics Research Lab, Stanford University, summer 2018 edition, 2018.

Julia Markovits. Acting for the right reasons. *The Philosophical review*, 119(2):201–242, 2010.

Christopher Noonan. Evolutionary debunking arguments, explanationism and counterexamples to modal security. *Erkenntnis. An International Journal of Analytic Philosophy*, June 2023.

Robert Nozick. *Philosophical Explanations*. Philosophical Explanations. Belknap Press, Cambridge, MA, January 1981.

David Owens. *Causes and Coincidences*. Cambridge University Press, January 1992.

Douglas W Portmore. Moral worth and our ultimate moral concerns. *Oxford Studies in Normative Ethics*.

Jonathan Schaffer. On what grounds what. In David Manley, David J Chalmers, and Ryan Wasserman, editors, *Metametaphysics: New Essays on the Foundations of Ontology*. OUP, January 2009.

Theodore Sider. *The Tools of Metaphysics and the Metaphysics of Science*. Oxford University Press, January 2020.

Keshav Singh. Moral worth, credit, and non-accidentality. In Mark Timmons, editor, *Oxford Studies in Normative Ethics, Vol. 10*. Oxford University Press, 2020.

Knut Olav Skarsaune. Darwin and moral realism: survival of the iffiest. *Philosophical studies*, 152(2):229–243, 2011.

Paulina Sliwa. Moral worth and moral knowledge. *Philosophy and phenomenological research*, 93(2):393–418, September 2016.

Michael Smith. *The Moral Problem*. Blackwell, 1994.

Philip Stratton-Lake. *Kant, Duty and Moral Worth*. Routledge, New York, 2000.

Sharon Street. A darwinian dilemma for realist theories of value. *Philosophical studies*, 127(1):109–166, 2006.

Sharon Street. Objectivity and truth: You'd better rethink it. *Oxford studies in metaethics*, 11(1):293–334, 2016.

Erik J Wielenberg. On the evolutionary debunking of morality. *Ethics*, 120(3):441–464, April 2010.