Trying to Act Rightly

by
Zoë A. Johnson King

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in The University of Michigan
2018

Doctoral Committee:

Professor Brian Weatherson, Chair
Professor Sarah Buss
Professor Scott Hershovitz
Associate Professor Maria Lasonen-Aarnio, University of Helsinki

Zoë A. Johnson King

zoejk@umich.edu

ORCiD: 0000-0001-5248-472X

Acknowledgements

This has been a long time coming. I am more grateful than I'll ever be able to properly put into words to an enormous number of people who've helped along the way.

I am most grateful to my Mum and Dad. I showed the early warning signs of an interest in Philosophy from a young age, and, amazingly, my parents both encouraged me rather than trying to get me to be more normal. My Mum has a story of a time in my childhood when we were having an argument and I shouted, "The trouble with you, Mum, is that *you're an assumer!!*" – which she apparently found amusing. Neither of us remember what that argument was about. But I've been questioning assumptions ever since. With all the patience of a saint, my Mum has tolerated each skeptical eyebrow and every excessively precise distinction, occasionally throwing back a distinction or useful concept of her own. I can't believe she's put up with me for so long. My Dad, meanwhile, has always gently encouraged both my adventurous streak and my rebellious streak. I don't think there's any chance that I would have quit my respectable job to move 3,000 miles away and do a PhD at a school that nobody back home has heard of if it wasn't for the long-term effects of this early conditioning. Nor would I have had the courage, drive, and/or whimsy to do almost any of the other stuff that I've tried over the years. At some point I plan to get the phrase *"and miles to go before I sleep"* tattooed around each ankle as a kinda homage to my parents. Thank you for continually reminding me that I am Saturday's child.

I first got the bright idea that I would go to University and study Philosophy at the tender age of 11, when I found a book called *The Philosophy Files* that a clueless librarian had left among the teenage fiction at my local library. I was intrigued by this weird green book with purple writing and a big pink pig on the front, and even more intrigued by the ideas inside – each chapter is dedicated to a philosophical question, starting with "Should I eat meat?" and ranging through "How do I know the world isn't virtual?" and "Where do right and wrong come from?" to "Does God exist?". I was hooked. I became a vegetarian, decided I was also an error theorist (and didn't notice the tension there for another seven years), and got pieces of lined paper from my Mum to write down what I thought about the questions in each of the chapters, figuring that this would be a good way to fix the thoughts swimming round my head. *The Philosophy Files* is by Stephen Law, who I have never met and who doesn't know any of this. He is going to receive a "your book changed my life" letter from me when I graduate in August. I hope he likes it.

The school that I was attending at that time, and the one that I then moved to in order to escape bullying, were both not-so-great academically. (The bit about deliberately under-performing in tests in "Accidentally Doing the Right Thing" is completely true!) But, at Wheldon, I was lucky enough to fall in with a group of ten defiant misfits whose strength in numbers meant that everyone pretty much left us alone to do our own weird thing – including the teachers as well as the other students. Nicola, Megan, Louise, Katie, Tilly, Haley, Bex, Lauren, and Sara, you gave me strength. Thank you. R.I.P., Haley.

I had idly carried the belief that I was going to go to University and study Philosophy in my head for several years when I had a meeting with a careers advisor that solidified this idle belief into a real intention. An organization called "Connexions" is employed by the

U.K. government to go into schools in low-income areas and talk to each student to try to make sure that they have some kind of plan for what to do when they finish school, and that it's more realistic than "be a footballer" or "be a pop star". I told my careers advisor that I wanted to go to University and study Philosophy. After a long pause, he said, "Huh. Have you considered hairdressing?", and gave me a pamphlet about a vocational course in hairdressing. I remember this meeting vividly and think about it often. I know that you will never read this Dissertation, Nick from Connexions, but I am so furiously happy to have proved you wrong. Thank you for making me determined to do this. There are tears in my eyes as I write this paragraph.

I went to a different school for the last two years before Uni. Bilborough College is a truly amazing place – it's still a non-fee-paying school, but it gets good results and you have to apply to get in, so it's basically full of all the clever non-wealthy seventeen- and eighteen-year-olds from Nottingham and the surrounding area. I went to Bilborough because it offered an A-level in Religious Studies, Ethics and Philosophy. The teacher who ran this course, Jake Dartington, also ran a Philosophy reading group on Monday lunchtimes. I remember going to the group's first meeting in week 1 of my first year, and being stunned by the fact that there were around twenty second-years in the room with me, they had all read Book I of the *Nicomachean Ethics*, and one of them was giving a presentation on virtue and happiness while the others asked questions. It was abundantly clear that everyone in the room genuinely cared about thinking through the issues. I had never met anybody else my age who liked (or had heard of) Philosophy before. I cried silent tears of joy, was ashamed, and tried to hide my tears behind my handout. In my R.S. class, I then met three other people who all also wanted to go to University and study Philosophy. We were so keen that Jake decided to introduce another A-level in straight Philosophy in our second year, using us as a pilot group. Anna, Sib, and Rossi, you were the best classmates ever –

not only did we make genuine philosophical progress in our conversations together every day, but you also helped me to feel like I wasn't crazy or foolish to want to spend three years just thinking deeply about stuff. *Thank you*. Jake, you were the best possible mentor. I owe you a debt that I'll never be able to fully repay. Thank you for your insights, your infectious calm, and your gentle encouragement. When I later met Richard Dawkins in a hotel lobby in Pennsylvania I wanted to sock him one for ya, but then I remembered that you probably wouldn't want that because Buddhists are usually not big fans of violence.

Bilborough had a system where teachers who thought that one of their students had "got what it takes" could anonymously refer them for the Oxbridge application scheme, which prepared us to apply to Oxford or Cambridge with helpful basics like explaining what on earth is going on with the interviews, telling us when the open days are, and talking us through the application form. Someone anonymously referred me for this scheme. There is absolutely no way that I would have applied to Cambridge, and no way that I'd be where I am now, if they hadn't. So, I owe this person more-or-less everything I have. I don't know who it was, but I'm guessing/hoping it was Jake.

I very nearly didn't get in to Cambridge. I applied to Emmanuel, and was one of twenty-five applicants each having two interviews, doing a logic test, and writing a timed essay for what turned out to be one spot. I didn't get the spot, but they liked me, so they kept my application file for other colleges to look at. I figured that I wasn't getting in, and was looking forward to going to Sheffield – which would, I'm sure, have been great! – when I got a call from my Mum while I was on the bus one morning saying that she'd found a message on our answerphone from Fitzwilliam College asking why I hadn't replied to their email and whether I was coming to my interview tomorrow morning. I frantically

bought train tickets and filled out paperwork to get the day off college. I am grateful to my Mum for checking our answerphone that day, and grateful to everyone at Bilborough who helped me to make it to the interview.

When I got to Fitz I met Michael "MDP" Potter. He kicked things off in characteristically terrifying style by leading me to his study, leaning back in a big ornate chair, and saying, "So, you want to come to Cambridge and study Philosophy. Do you realize that, if I let you in, the government will subsidize thousands of pounds toward your education over the next three years? You do? Well, then, *justify that*." Thus Dr. Potter started very much as he meant to go on: imposing, cold, mildly pleased when I caught on to things quickly but slightly impatient when I gave silly answers. I am sure that I ultimately owe more of my intellectual development to this person than to anyone else. He corrected my essays in painstaking detail, including the many spelling, grammar, and punctuation errors with which they were littered as well as the philosophical content, and he minced no words in telling me exactly what he thought of my ideas, as well as of everyone else's. He also set me up with an outstanding array of supervisors, from whom I learnt more over the course of nine eight-week semesters than I would ever have thought possible. To the extent that I turned from coal to diamonds over my undergraduate degree, it was Dr. Potter who did the polishing. He set such high standards and seemed so difficult to impress that I was certain I'd fail my exams and get kicked out in first year, when in fact I ended up coming top of the year. Precisely this then happened again in my second and third years. I have Michael to thank for all of my awards and prizes, for the philosophical skill by means of which I earned them, and possibly also for the nagging sense that I will never be good enough. I recently learned from a friend at Cambridge that Michael had asked after me, and I was delighted to discover that he still remembers who I am. Apparently he thought highly of me all along.

All of my supervisors were amazing. Looking back on it now, and knowing more about how other undergraduate degrees are structured, I can hardly believe that I would write a two-to-six-thousand-word essay and then an expert in the field would read it and talk to me about it, *every week*. I am so grateful to all of them for engaging with my half-baked ideas. That said, there were two supervisors who had a particularly enormous formative impact on both my philosophical thinking and my burgeoning self-esteem, so they each deserve special mention here. Ben Colburn helped me through repeated existential crises brought on by excessive metaethics in some cases, and by an overdose of worrying about injustice in others. I'm sure he could tell that this was no mere intellectual exercise for me and that I was genuinely very bothered about things like whether anything is wrong and what it's okay to do to people. (This attitude still strikes me as quite reasonable.) We went on walks for some of our supervisions, often with dramatic clouds mirroring the drama of the conversation. Ben was sharp, kind, and clearly also genuinely concerned about the answers to questions in value theory, so I was deeply invested in what he thought of my work; I recall my nerves steadily rising in the time leading up to each supervision, which he would begin either by saying "I liked this essay" or "I was disappointed by this essay", and my mood would soar or plummet accordingly. Meanwhile, I learnt how to produce extended pieces of proper philosophical writing – that is, the sort that you don't just type out in a manic all-nighter, but that you develop slowly based on a careful survey of the literature and then revise repeatedly over several weeks – from Hallvard Lillehammer. Hallvard was a fantastic supervisor. My second-year essays, third-year dissertation, and M.Phil thesis would have sucked if it wasn't for his judicious and patient guidance. The care with which Hallvard read my work, and the savvy tact with which he then dispensed advice for revisions and comments about the general arc of the project, set a high bar for my professors in graduate school. I also remember Hallvard once comforting me during

a meeting when I was very distressed because I was afraid of one of my examiners. I am grateful for his rare combination of intellectual rigor and emotional intelligence.

At Cambridge I met Will King, with whom I was in a relationship for six years. Will did more to calm my nerves during that time than anyone else has, and he was a great person to be around – he was unusually normal and unusually emotionally well-adjusted for a Cambridge undergrad, and was always just the right mix of sincere and playful. We went through a lot together and went on a lot of fun adventures. Will was maximally supportive of me in everything that I ever wanted to do, even up to and including the move to the USA that ultimately ended our relationship. I am still deeply grateful to him for all of this.

After my M.Phil, I left academia. This was partly because I wanted to do something more "difference-makey", and partly because I had been routinely overwhelmed at Cambridge by the fact that my new friends from Uni had such different etiquette and so much more favorable employment opportunities than my old friends from school, and yet they didn't seem all that much smarter to me. I was enchanted by the Teach First programme's refrain that "no child's educational potential should be determined by their parents' economic background" – I thought this was exactly right, and a great way of putting it. So, I joined Teach First. This was by far and away the most challenging and the most rewarding thing that I have ever done. I am immensely grateful to all my instructors, LDOs, and peers on the programme, not only for teaching me how to teach, but also for making me care a *lot* about teaching, and to recognize it for the liberatory tool that it is. I am also immeasurably grateful to my friend and mentor Laura Shuttleworth for sharing a million tips, a million jokes, and a million uplifting words when someone (usually our awful boss!) was being

an arsehole, over the years. Laura saw how passionate I was about helping our students, and she in turn helped me to navigate through all the bureaucracy, paperwork, and dumb institutional politics that got in the way. She also taught me how important it is to look after yourself in teaching, and not to let thoughts about how much more good you could do for Kid X if you just keep working on Thing Y lead you to push yourself to the point of total mental, physical, and emotional exhaustion. There's a song by Bat for Lashes that goes, "Laura, you're more than a superstar" – when I hear that song, I think of this Laura.

More than anyone else from this period of my life, I am grateful to my students at Quest Academy for making the early mornings, late nights, and exhaustion all *so* worthwhile – for being hilarious, curious, quick-witted, cheeky, frustrating, endearing, and vulnerable, and for slowly coming to trust me, work in my classes, see themselves making progress and take pride in it. I am grateful for every time they pushed themselves to produce work that exceeded all expectations – like the time when Shaniya Robinson wrote a play called "God on Trial" about the arguments for and against theism, or the time when our local MP visited the school and Eve Abiodun wrote a hard-hitting cartoon that challenged him about the inadequacy of the Millennium Development Goals, which she had to deliver in cartoon form because she was doing a vow of silence for charity. I am also grateful for the spirit of open-minded, sincere inquiry with which my students approached topics in class – like the time when Ayo Lasisi, exacerbated by a thought-experiment, pushed his chair back from the table and exclaimed "No, miss! You don't understand! *I don't know what's right and what's wrong!!*" – and I am grateful for the playful spirit with which they approached school in general – like the time at the open evening in my first year (which I was quite nervous about) when Brandon Kandisai found a pair of giant golden pants in the storecupboard and proceeded to wear them all evening. Most of all, I am so grateful for, and so proud of, the ambition that I managed to foster in my students despite their

adverse circumstances – like when I helped Dilfaraz Keftani to become a prefect despite having arrived from Afghanistan with no English just three years previously, or when Octavia Fairbrass decided aged thirteen that she wanted to be a politician and I helped her to write to our MP to ask for an internship. Lastly, I am grateful to all the students who, when I told them that they should go to Uni because they were so smart that they could even do a PhD someday, cheekily replied that I was smart and I should go and do my PhD. Eventually, I realized that they had a point. I don't think I'd be here if it wasn't for them, and I certainly wouldn't be the sort of person I am today without these many thousands of uplifting moments in my past.

The rest of the credit for this Dissertation goes to people I've met over the past five years at Michigan. There are too many to list them all. But I'll try to list some of the main ones. I am grateful to all of the professors whose classes I took – Jim Joyce, Ishani Maitra, Allan Gibbard, Rich Thomason, Louis Loeb, Victor Caston, and Sarah Moss – for teaching me their craft. I am grateful to all of the organizers and presenters at all the conferences and workshops and summer schools that I've attended, and to everyone who has ever asked a question or made a comment at one of my talks, for helping me to refine my ideas. I am grateful to Ralph Wedgwood for arranging my visit to USC and for making so much time to talk while I was there, to Mark Schroeder and Steve Finlay for finding time to chat too, and to Jake Ross and Robin Jeshion for welcoming me into their classes. I am immensely grateful to Nomy Arpaly for graciously taking me under her wing, keeping her talons at bay throughout our meetings, dispensing a great deal of wisdom, and deciding that I am an egret. And I am grateful to a smorgasbord of wonderful graduate students at Michigan and beyond. Boris, Kevin, Chris: thanks for being my partners in crime in so many classes, helping me to feel confident speaking up, and for developing our intellectual sensibilities together. Sydney, Sara, Cat, Daniel: thanks for showing me how to grad school, and for

helping me to understand how academic Philosophy works. Eli, Johann, Kevin-Plaid, and Alvaro: thanks for being the most eclectic and eccentric mix of philosopher-roommates a girl could dream of. Robin, Jon, Caroline: thank you for helping me to build our outreach program from teeny-tiny beginnings into the awesomeness it is today. Dmitri: *Oehor cro oru suyehmomon gr fhoc oemong or fs, ohermon gr fhoc oemong oeurone lmoe fs, hoe eseymon fs or es rhefsu. M'f gruuc mo emeo'o lrur roo*. Rima, Nathan, Renee, Alex, Nicola, Kirun, Ian, Mary, YongMing, Zach, and all the rest of the gangs at USC and Brown: thank you all so much for welcoming me into the fold, and for showing me how you do things out there. Maegan: thanks for being the most surprisingly perfect market buddy I could have had, and thanks a million for all the emotional support and pangolin gifs long before it became clear that we were just gonna swap. Ginger and Antonia: thank you both for so, so much validation and encouragement – you kept me sane during the first three months of this year. Elise and Carolina: you are the best. Thank you for all the jokes, gifs, memes, uplifting messages, celebrations of the good and commiserations for the bad, and all-round top-quality life coaching. And thanks for the legendary peach emoji earrings.

Lastly, I am endlessly grateful to my committee members. I know that every grad student thinks that they have the best committee. But they're all wrong, because I do. I am forever grateful to Maria Lasonen-Aarnio for believing in me and telling me so, and also for being seriously badass while junior and female – as a naïve and impressionable young second-year, this was the proof of concept that I needed. I am forever grateful to Scott Hershovitz for being single-handedly responsible for my secondary research project, by pushing me to really make something out of what I thought was just a term paper. Maria and Scott have both provided me with a maximally productive combination of insightful feedback and steady encouragement over the years, and my papers and self-esteem are a great deal better for their influence on them. I have also been unbelievably fortunate in having had

the chance to work closely with Brian Weatherson and Sarah Buss. A better combination of mentors for me does not exist. I am so much more confident now than my undergrad self would have thought possible, and this is mostly due to Sarah's unfettered enthusiasm for my work and Brian's insistence on talking to me as he would to any peer or colleague. I have benefited enormously from Brian's encyclopedic knowledge of the discipline, both in terms of its intellectual history (and geography!) and its bizarre internal politics, and from his willingness to share his shrewd reflections on all these fronts. My work has also benefited beyond measure from Brian's masterful command of logical space, and from his consequent ability to immediately discern how things I say might commit me to fun or contentious or surprising other things further down the line, as well as precisely how the things I say compare to everything else previously said by anybody. My advice to junior graduate students is always to find the sharpest and least egotistical person who thinks the opposite thing to you and to work with them, which is a reflection on my work with Brian. Meanwhile, my conversations with Sarah have always left me simultaneously clearer about what really matters to me and more confident in my ability to express it, both by several orders of magnitude every time. Sarah has a miraculous ability to discern what I really think when I'm just saying some hand-wavey gobbledegook, and to kindly tell me what it is. Her questions for me are always maximally probing, and she always fast-tracks our conversations to the heart of the matter. It is largely thanks to Sarah that I have a "big picture" research project and a sense of my identity as a scholar, rather than just a bunch of ideas.

Thanks, everyone. I couldn't have done it without you.

Abstract

My research focuses on the moral evaluation of people's motivations. A popular recent view in Philosophy is that good people are motivated by the considerations that make actions morally right – the "right-making features". For example, this view entails that a Black Lives Matter protester can be a good person if she is motivated to engage in protest by the thought that it will bring about *equality*, or *justice*, since this is what makes engaging in protest morally right. But this view entails that the protester cannot be a good person if she engages in protest *because it is morally right*. I think that this is a mistake. My view is that it is good to be explicitly committed to acting rightly and motivated by the moral rightness of one's actions.

More specifically, I explore the nature and defend the value of a complex state that I call *trying to act rightly.* This comprises (a) wanting to act rightly, (b) thinking about which actions are right, and (c) doing the things that you think are right, because they are right. The three papers of my Dissertation each make part of the case for trying to act rightly.

My first paper, "Praiseworthy Motivations", addresses the view that it is good to be motivated by the right-making features but not good to be motivated to act rightly. I argue that this view rests on poorly-drawn comparison cases that are not genuine minimal pairs, and that well-constructed cases show these two types of motivation to be equally good. I address the worry that trying to act rightly leads people with false moral beliefs to act wrongly, by noting that this also applies to motivation by right-making

features, since people can be motivated by a right-making feature while being mistaken about which acts have this feature. I then argue that we should distinguish carefully between motivations, actions, and beliefs when evaluating these well-meaning but morally mistaken agents.

The second paper, "We Can Have Our Buck and Pass It, Too", addresses the view that the fact that an act is morally right is not a genuine reason to perform it, and that our reasons for action are instead provided by the right-making features. I argue that this view rests on a confused picture of moral metaphysics, which would rule out any case in which one reason to perform an act is partially metaphysically constituted by another fact that is also a reason to perform the same act – as, for example, when a salad both *is healthy* and *contains vegetables*. I then sketch an alternative picture of moral metaphysics, on which genuine reasons for action can be metaphysically related to one another.

My third paper, "Accidentally Doing the Right Thing", uses general reflections on the nature of deliberate action and its relationship to praiseworthiness to argue that someone is only praiseworthy for acting rightly if she was trying to act rightly. I apply this idea to the philosophical debate on moral worth, defending the Kantian view that actions have moral worth just in case they are instances of someone's trying to act rightly and succeeding. This is a radical departure from the most popular contemporary view on moral worth, and requires a re-evaluation of the main case discussed in this literature – that of Huckleberry Finn.

Table of Contents

Introduction

My dissertation explores the nature and defends the value of explicitly moral motivation.

This is the kind of motivation that someone has when she faces a complex, fraught, highly morally charged situation, and thinks to herself something like, "Sheesh, I really want to do the right thing here – I just wish I knew what that was!" In this case the agent's motivation is explicitly moral: she wants to do *the right thing*, whatever it may turn out to be, and she thinks of what she wants to do in these explicitly moral terms.

Someone can also exhibit explicitly moral motivation when she thinks that she knows what the right thing to do is (or, at least, when she takes herself to have a reasonably good guess). Sometimes, someone takes herself to have established what is morally required of her under her circumstances, and she does this thing *because it's the right thing to do*. For example, someone might choose to engage in a political protest despite knowing that this threatens her prudential interests by posing a risk to her personal safety, and might be moved by her conviction that participating in this protest is morally right. In this way, an agent's explicitly moral concern can propel her to action.

More specifically, I am interested in a state that I call *trying to act rightly*. This is a complex state that someone can be in only over a period of time, which has three main components:

a) Wanting, intrinsically, to do the right thing – i.e., to do whatever is in fact morally required of you under your circumstances.
b) Engaging in moral inquiry aimed at figuring out what is morally required of you, with a view to then doing it.
c) Doing the things that you take to be morally required of you, and, in each case, doing the thing *because it's the right thing to do*.

Here is a quick bit of clarification. The desire in (a) can be a specific desire pertaining to a particular set of circumstances (as in the example of the morally uncertain agent), or it can be the result of a general desire to do whatever is morally required, applied to a particular set of circumstances. What is important is that the desire is *intrinsic*. This means that the agent wants to do the right thing just because it's right, rather than because it's right and she'll be financially rewarded for doing what's right, or because it's right and someone that she finds attractive will go on a date with her iff she does what's right, or etc.

It is my view that trying to act rightly, so construed, is good. That is to say: I think that trying to act rightly is at least part of one way of being a good person.

This view may seem obvious or trivial. It is neither. On the contrary, very many ethicists and metaethicists have denied the view that I hold. Those who deny this view typically do so as a result of their accepting a certain supposed distinction between two types of moral motivation. I think that this distinction is mistaken and misleading. But, for the purposes of introducing my research project, I will discuss it here.

Some philosophers think that we can usefully contrast two types of moral motivation using the *de re*/*de dicto* distinction from philosophy of language. These philosophers use

the phrase "motivation by rightness *de dicto*" to refer to the kind of explicitly moral motivation that I defend. This is a motivation with the concept of moral rightness, or a cognate concept, as part of its content, hence the idea that we are talking about rightness *de dicto*. And these philosophers use the phrase "motivation by rightness *de re*" to refer to any motivation that has as its object one of the features that *make* acts right, according to the true moral theory – collectively called "the right-making features".

What counts as motivation by rightness *de re*, so construed, depends on which moral theory is true, as it depends on what the right-making features are. In the literatures with which I am primarily engaged, we typically try to remain as neutral as possible on this question. When we need examples, we help ourselves to plausible-seeming assumptions about the sorts of things that might be right-making: we assume that the right-making features include considerations pertaining to well-being, fairness, equality, honesty, justice, and other things that pre-theoretically seem morally significant. I follow my philosophical opponents in this regard (though I think that there is a risk of theoretical sloppiness here, which I explain in footnote 1).

Those who accept the contrast between motivation by rightness *de dicto* and *de re*, as just described, typically do so in order to defend a popular combination of evaluations of these two types of motivation: they venerate motivation by rightness *de re*, and they denigrate motivation by rightness *de dicto*.

The most famous statement of this combination of evaluations comes from Michael Smith, in his 1994 book *The Moral Problem*. There, in a much-read and widely cited passage, Smith famously offers the opinion that "commonsense tells us that being [motivated by rightness *de dicto*] is a fetish or moral vice, not [a] moral virtue" (p.75). Smith's position here is strong. According to him, not only is motivation by rightness *de*

dicto no part of any way of being a good person, it is a bad thing – a "vice", and thus something that counts against the agent in an assessment of her character. But, although this view is strong, it has plenty of adherents. Many metaethicists have reported sharing Smith's so-called "fetishism intuition" (e.g. Miller 1996, Copp 1997, Dreier 2000, Zangwill 2003, Toppinen 2004, Strandberg 2007). Smith himself takes it to be closely related to Bernard Williams' "one thought too many" intuition (Williams 1981, p.18; see Smith 194, pp.76-77), which is also popular. And some philosophers writing in distinct but related literatures have simply assumed that Smith's intuition is correct; for instance, Brian Weatherson (2014) assumes that Smith is correct in his work on moral uncertainty.

Some philosophers prefer a weaker view, according to which motivation by rightness *de dicto* is neither a virtue nor a vice. On this weaker view, the criterion for identifying good motivations is straightforward: the good motivations are all and only the motivations that have right-making features as their objects. This entails that an explicit concern for acting rightly does not make the good list. That is because, no matter what the right-making features may turn out to be, rightness *itself* cannot be among them. Acts cannot be *made* right by their rightness itself; this would be circular. So, if the list of good motivations contains all and only motivations whose objects are right-making features, then motivation by rightness *de dicto* is not on the good list. On this weaker view, therefore, motivation by rightness *de dicto* is not a vice, but it is not a virtue either. So, on this view, the kind of motivation that I defend in this Dissertation is morally neutral – that is to say, it is morally on a par with a motivation to eat some hummus, or a motivation to go for a run, or a motivation to do some other morally innocuous thing.

The weaker view also has plenty of adherents. Some philosophers write book-length defenses of it, like Nomy Arpaly and Timothy Schroeder (2013). And, as was the case with Smith's stronger view, some philosophers writing in distinct but related literatures

have simply assumed that this view is correct. For instance, Julia Markovits (2010) assumes that this view is correct in her work on moral worth, Alison Hills (2009) assumes that this view is correct in her work on moral testimony, and David Shoemaker (2007) assumes that this view is correct in his work on membership of the moral community.

I think that both the strong and weak views just described are incorrect. In this Dissertation, I aim to refute them. Clearly, I disagree with both views on evaluative grounds: I think that trying to act rightly is good, whereas these views both entail that it is not. But this clash of evaluative intuitions may end in a stalemate, with neither side being able to shift the other's intuitions in their preferred direction. I avoid stalemate by criticizing a different part of both the strong and weak views: I challenge the supposed distinction between two types of moral motivation on which these views both rest. I maintain that, when we make more of a concerted effort to spell out the *nature* of these types of motivation – that is to say, when we get clearer on the psychology, the epistemology, the metaphysics, and the semantics – we reveal a picture on which they have far more in common than has traditionally been recognized, and indeed are roughly on a par.

The most important point to understand in this regard is this: *the right-making features are not fundamental*. No plausible candidate for being a right-making feature is such that facts about its instantiation are brute facts, insusceptible of further explanation. For instance, suppose that an act is made right by its being fair: it is morally right because it is fair. Its being fair is not then a brute fact, insusceptible of further explanation. Quite the contrary. If an act is made right by its being fair, then it is made fair, in turn, by further features of the act – perhaps that it distributes social benefits and burdens on reasonable, non-arbitrary grounds. We could call this a "right-making-feature-making feature", since it is a feature of an act that makes it the case that the act instantiates a right-making feature

(fairness). But this right-making-feature-making feature is also not fundamental. If an act distributes benefits and burdens on reasonable, non-arbitrary grounds, that in turn is made the case by further features of the act – perhaps that it is meritocratic, or that it makes reparations for past injustice, or that it gives resources to those with the highest need. And *these* features are not fundamental, either. So on we might go, mapping out a metaphysical hierarchy of features of acts that ranges from the less to the more fundamental, with rightness at the top and either fundamental moral facts or facts about the location and speed of physical particles at the bottom (depending on the truth of moral naturalism, which I will take no stand on here).[1]

I think that, once we zoom out like this and begin to see the entire metaphysical hierarchy, the idea that there is something special about the *right-making features* in particular – the features at level 2 – seems silly. There is nothing special about level 2. That is to say, there is nothing special about right-making features, as opposed to the right-making-feature-making features (and all other, less fundamental, moral features) below them and to moral rightness above them.

The second-most-important point to understand is this: the *de re*/*de dicto* distinction has been misapplied. Philosophers write and speak as if the distinction applies to attitudes themselves; as if there is such a thing as a *de re* motivation or a *de dicto* motivation. But this is not the case. The *de re*/*de dicto* distinction applies to our *ascriptions* of attitudes, not to the attitudes themselves. This means that there are such distinctions to be drawn with respect to any motivation whatsoever, including any motivation whose object is any

[1] Understanding that moral properties can be arranged in metaphysical hierarchies helps us to see why there is a risk of theoretical sloppiness in simply grasping for anything that pre-theoretically seems morally significant, when looking for examples of right-making features. Something's pre-theoretically seeming morally significant does not tell us where in the true metaphysical hierarchy it is located, so it does not tell us that the feature is a *right-making feature*, rather than something lower down. For instance, meritocracy may seem, pre-theoretically, to be morally significant, while being a *realizer* of a right-making feature (namely fairness) rather than a right-making feature itself.

feature in the metaphysical hierarchy just sketched. The distinction has nothing special to do with moral rightness.

For instance, take fairness. Someone could be motivated by fairness either *de dicto* or *de re*: she could have an explicit concern with acting fairly (*de dicto*), or a concern with whatever it is that falls immediately below fairness in the true metaphysical hierarchy (*de re*). The same holds of any other property in a metaphysical hierarchy. Someone could have a motivation whose object is that property – in which a concept referring to the property figures explicitly – which would be motivation by the property *de dicto*. Or she could have a motivation whose object is whatever falls directly below the property in the true metaphysical hierarchy, which would be motivation by the property *de re*. Or she could have both motivations at the same time.

These points help to show why the *de re*/*de dicto* distinction does not apply to attitudes themselves. Consider again an agent's explicit concern with acting fairly. Assuming that fairness is a right-making feature, this attitude can be described as motivation by rightness *de re*. But the same attitude can equally accurately be described as motivation by fairness *de dicto*. Those are just two ways to refer to a single motivation. The motivation itself is neither *de dicto* nor *de re*. It is just a motivation with an object: acting fairly. The *de dicto*/*de re* distinction applies to our ways of describing this motivation when we ascribe it to the agent.

Once we recognize this, it becomes clear that many supposed distinctions between motivation by rightness *de dicto* and *de re* are spurious, and that many criticisms of motivation by rightness *de dicto* apply with equal force to motivation by rightness *de re* – or, indeed, to any other motivation whose object is any feature in a metaphysical hierarchy. Criticisms of motivation by rightness *de dicto* often focus on cases involving

agents in unfortunate epistemic positions, who want to act rightly but are uncertain or mistaken about what the right thing to do is. These criticisms extend straightforwardly to motivation by rightness *de re*, as it is possible to be motivated by any right-making feature while being uncertain or mistaken about its precise nature and extension. For instance, someone may be explicitly concerned with acting fairly but uncertain or mistaken as to what fairness consists in, and thus may end up acting unfairly, though she was *trying* to act fairly. Indeed, the relevant phenomena are not even confined to moral properties; for example, someone may be motivated to eat a tomato while being uncertain or ignorant as to whether an item in front of her is a tomato, and thus may end up inadvertently eating a persimmon. In short, it is possible to be motivated by any feature whatsoever *de dicto* but not *de re*, if one is uncertain or ignorant about it. This means that any problematic phenomena in this vicinity are problems with motivation and action in general. They have nothing to do with trying to act *rightly* in particular.

That is the take-home point to glean from my first paper, "Praiseworthy Motivations". In this paper I argue that motivation by rightness *de dicto* and *de re* have been unfairly compared, using cases that are not genuine minimal pairs. The existing literature has focused on pairs of cases in which one agent is motivated to act rightly but has false beliefs as to what rightness consists in, thus ending up acting wrongly, while another agent is motivated to perform acts with some right-making feature and has true beliefs about what this feature consists in, thus ending up acting rightly. These are not minimal pairs, since they differ in multiple respects relevant to our moral assessment of the agents besides the key issue of whether they are motivated by rightness *de dicto* or *de re*. I then argue that, when we compare genuine minimal pairs, the idea that there is a substantial difference between the praiseworthiness of motivations whose objects are right-making features and a motivation to act rightly becomes difficult to sustain. I first consider good cases, in which the agents have true beliefs about moral metaphysics and are aware of

what it is that the property by which they are motivated consists in (and of that which that-which-the-property-consists-in consists in, and that which *that* property consists in, and so on). These agents succeed in doing what they are motivated to do, and thus act morally rightly. Intuitively, good cases like this are not rendered markedly worse by the agent's having been intrinsically motivated to act *rightly*, rather than being intrinsically motivated by one of the features at level 2 of the metaphysical hierarchy. I then turn to bad cases, arguing that the phenomena described above – the right-making features are not fundamental, and it is possible to be motivated by any features *de dicto* but not *de re* – ensure that there are direct analogues for motivation by rightness *de re* of all worries about agents in unfortunate epistemic positions trying to act rightly but ending up acting wrongly. I suggest that the appropriate response to these worries is to be more fine-grained in distinguishing things for which an agent might be praiseworthy, recognizing that people are not either wholly perfect or wholly awful, and acknowledging that someone may be praiseworthy in many ways while still falling short (or being positively blameworthy) in other ways. I then use the phenomena described above to develop an approach to evaluating motivations, the "partial credit" approach, which enables us to recognize the precise extent of each agent's moral success.

My second paper, "We Can Have our Buck and Pass It, Too", develops a picture of moral metaphysics that fleshes out some of the metaphysical claims in the first paper and thereby helps to avoid another family of wrong-headed criticisms of motivation by rightness *de dicto*. The family of wrong-headed criticisms at issue centers around the claim that the fact that an act is morally right is not a reason to perform it, and our reasons for action are instead facts about acts' right-making features. Philosophers engaged in the denigration of motivation by rightness *de dicto* and veneration of motivation by rightness *de re* often help themselves to this claim about reasons (see e.g. Markovits 2010, pp.207). And this claim is suggested by a currently-fashionable move in metanormative theory,

called *buck-passing*, applied to moral rightness. In general, buck-passers about a moral property *M* make two claims about *M*: that the instantiation of *M* is not itself a reason for anybody to do anything, and that *M* is rather a positive status that things have in virtue of our non-*M* reasons for actions or attitudes. Buck-passers defend their approach by appeal to an intuition that I call "the redundancy intuition", which holds that it is redundant – or, worse, an illegitimate form of double-counting – to say that facts about moral properties are reasons, having acknowledged that facts about the *M*-making features are already reasons. Buck-passing is popular but controversial (for defenses see e.g. Scanlon 1998, Parfit 2001, Olson 2004, Suikkanen 2004, Stratton-Lake and Hooker 2006, Skorupski 2007, and for criticisms see e.g. Rabinowicz and Rønnow-Rasmussen 2004, Crisp 2005, Väyrynen 2006, Liao 2009, Gregory 2014). I argue against buck-passing about moral rightness. On my view, the fact that an act is morally right is a genuine objective normative reason to perform it, and that the fact of its possessing some right-making feature is *also* a genuine objective normative reason to perform it. I argue that the redundancy intuition cannot be probative in telling us which facts are reasons and which are not, since it massively overgeneralizes: it applies to *any* case in which one fact that seems to count in favor of performing a certain act or adopting a certain attitude is metaphysically constituted, partly or wholly, by another fact that seems to count in favor of performing the same act or adopting the same attitude. This is true of many plausible candidate right-making features just as much as it is true of rightness. And it is even true of many non-moral features of acts. I offer a way out of the mess by suggesting that relationships of metaphysical constitution can obtain between genuine objective normative reasons, and suggesting a way of thinking about moral metaphysics – the "share the weight" view – that explains how this is possible.

These first two papers defend positions that are permissive. I do not deny that motivation by rightness *de re* is praiseworthy, nor do I deny that facts about right-making features

are objective normative reasons. Rather, I argue that motivation by rightness *de dicto* is *also* praiseworthy, and that the fact that an act is morally right is *also* a genuine objective normative reason. So, while these papers do develop novel positive proposals for assessing agents' motivations and for conceiving of moral metaphysics (the partial credit approach and the share the weight view), in the dialectic between defenders of motivation by rightness *de dicto* and *de re* their role is largely defensive. The papers show that some prominent criticisms of motivation by rightness *de dicto* either apply with equal force to motivation by rightness *de re*, or do not apply to anything because they are fundamentally confused. The third and final paper of this Dissertation – "Accidentally Doing the Right Thing" – is different. In this paper I go on the offensive: I discuss a genuine difference between motivation by rightness *de dicto* and *de re*, as traditionally construed, but one that I think redounds to the credit of motivation by rightness *de dicto*. I argue that someone's trying to act rightly is necessary for her to count as *deliberately* doing the right thing, which in turn is necessary both for being praiseworthy for doing the right thing and for performing acts with genuine moral worth. I make this case using some general reflections on the nature of deliberate action and its relationship to praiseworthiness. *Contra* recent philosophers who argue that it is sufficient for moral worth that an agent is motivated to do the right thing by its right-making features, I note that it is a central part of the concept of moral worth that an act lacks moral worth if it is an instance of someone's *accidentally* doing the right thing, and I then suggest that ordinary-language intuitions about doing things accidentally suggest that one may accidentally perform an act of type *T* even if one was motivated by the very features that make it the case that one's act is of type *T*. If someone has no idea that she is acting rightly, I argue, then she accidentally does the right thing – and her act lacks moral worth – even if she was motivated by the right-making features. I then argue that an agent is praiseworthy for performing an act of a certain good type only if she did so deliberately. If I am correct about this, then there are two positive moral properties (moral worth and

being praiseworthy for acting rightly) that an act or agent can have *only* if the agent exhibits the kind of explicitly moral motivation that I favor. This third paper thus defends a version of the Kantian view that an act has moral worth only if its agent does it because it's the right thing to do. On my view, the best interpretation of the concept of moral worth that we inherit from Kant construes the performance of an act with moral worth as a certain kind of achievement: the achievement of someone's trying to act rightly and succeeding.

The remainder of this Dissertation consists of the three papers just sketched, and then a concluding chapter summarizing the directions in which I would like to take this project in future work.

I: Praiseworthy Motivations

I'm just a soul whose intentions are good;
Oh Lord, please don't let me be misunderstood.
— Nina Simone

1. Introduction

In this paper I defend the following thesis:

> SYMMETRY THESIS: If motivation by rightness *de re* is praiseworthy, then so is motivation by rightness *de dicto*.

My thesis is equivalent to the negation of a popular view: that motivation by rightness *de re* is praiseworthy, but motivation by rightness *de dicto* is not. I will call this "the asymmetry thesis".

Let me begin by providing some context, which will explain what the symmetry and asymmetry theses are talking about.

We all face morally difficult decisions. Life is complicated, lots of things are morally significant, and it is frequently hard to tell precisely what is morally required of us.

Some people approach morally difficult decisions thinking something like "I just want to do the right thing in this situation, whatever it is!" These people then engage in moral reflection. When they think that they have worked out what the right thing to do is in their situation – or, at least, when they have a good guess – they then do it, *because it's the right thing to do*.

Other people have more concrete concerns. Faced with morally difficult decisions, they think about what would be *honest*, or *kind*, or *fair*, or about what's in the *interests* of the people concerned, rather than thinking about what's morally right *per se*. These people then choose a course of action based on its having one of these more concrete features, rather than choosing it based on its moral rightness.

But some of the more concrete features by which these people are motivated are among the features that *make* courses of action morally right – the so-called "right-making features". So, although people moved by these concerns are not motivated by the moral rightness of their actions *per se*, they are motivated by the very features that their actions' moral rightness consists in.

Philosophers distinguish between these two types of moral concern. We say that the first type of person is motivated by rightness *de dicto.* This means that she is explicitly concerned with acting morally rightly: she has a motivation with the concept of moral rightness, or a cognate concept, as part of its content. We say that the second type of person is motivated by rightness *de re*. This means that she is concerned with the features of actions that their moral rightness in fact consists in, according to the true moral theory (whatever it may be). The content of her motivation includes these "right-making" features, but it need not include the concept of moral rightness itself.

This way of drawing the distinction comes from Michael Smith's discussion of praiseworthy motivations in *The Moral Problem* (1994). Smith drew the distinction in order to denigrate motivation by rightness *de dicto*. He denied that this type of moral concern can be part of what it is to be a good person, claiming that "good people care non-derivatively about honesty, the weal and woe of their children and friends, the well-being of their fellows, people getting what they deserve, justice, equality, and the like, not… doing what they believe to be right, where this is read *de dicto* and not *de re.* Indeed, commonsense tells us that being so motivated is a fetish or moral vice, not [a] moral virtue" (*ibid.,* p.75). Smith envisages the "moral fetishist" as someone who is left cold by all mention of the honesty, justice, equality, etc., of an act, springing into action only when it is added that the act is *right*.

A lively debate ensued as to whether Smith is right about this. Some (e.g. Lillehammer 1996, Svavarsdóttir 1999, Olson 2002, Aboodi 2016) argued that Smith's so-called "commonsense" intuition is mistaken or misleading. Others (e.g. Miller 1996, Copp 1997, Dreier 2000, Zangwill 2003, Toppinen 2004, Strandberg 2007) reported sharing it. And Smith himself later clarified that what he really thinks is praiseworthy is not motivation by rightness *de re*, but rather an "executive virtue" by which agents' intrinsic motivations reliably track their beliefs about what moral rightness consists in (Smith 1996, pp.176-177).

This literature is already crowded with disputants, and I will not wade into it here.

I am interested in a different literature. Independently of the metaethical arguments for and against Smith's position, the asymmetry thesis – that motivation by rightness *de re* is

praiseworthy, but motivation by rightness *de dicto* is not – has become popular in its own right.

The most developed defense of the asymmetry thesis is from Nomy Arpaly and Timothy Schroeder (2013). Their view is that good will is a matter of intrinsically desiring that which is in fact right or good, *de re*, and/or *not* intrinsically desiring that which is in fact wrong or bad, *de re*, and that good will in turn is both necessary and sufficient for both virtue and praiseworthiness. They say that "it is the right or good conceptualized in the way preferred by the correct normative theory, and not merely via the concept RIGHT or GOOD, that motivates people moved by good will" (*ibid*., p.177). Thus they explicitly consider and reject the possibility that it might be praiseworthy to be motivated by the right or good *de dicto*. On this view, what matters for good will, virtue, and praiseworthiness is that an agent is motivated by the very features that rightness or goodness in fact consists in, and that she conceptualizes these features "in the way preferred by the correct normative theory".

Arpaly and Schroeder argue for their view by comparing agents, all of whom accept false and pernicious moral theories (such as a pro-slavery moral theory). Some are motivated by rightness *de dicto* but not *de re*, while others are motivated by rightness *de re* but not *de dicto*. Those who are motivated by rightness *de dicto* do what is in fact wrong, believing it to be right, since it is right according to their false moral theory. And those who are motivated by rightness *de re* do what is in fact right, believing it to be wrong, but being undeterred by this since they are uninterested in rightness *de dicto*. Arpaly and Schroeder emphasize that the latter (*de re* morally motivated) agents, who do what is in fact right and care about what is in fact morally significant, seem more praiseworthy than the former. They do not use the term "fetishist" to describe the former (*de dicto* morally motivated) agents, but their intuition here is roughly the same as Smith's. I will object to

this way of comparing cases in §3; for now, I simply note that the asymmetry thesis has received some sophisticated recent defenses.

Other authors, writing on related topics, have simply assumed that the asymmetry thesis is correct. For example, Brian Weatherson (2014, pp.152-154) deploys Smith's fetishism intuition as the key move in his argument against "moral hedging", which involves taking account of one's credences in various different moral theories when deciding what to do. Weatherson argues that someone would only engage in moral hedging if she were motivated by rightness *de dicto*. Then he suggests that this shows moral hedging to be objectionable, as it "is not possible without falling into the bad kind of moral fetishism that Smith rightly decries" (*ibid.*, p.154). Weatherson is explicit about the fact that this is his main argument against moral hedging.

Similarly, Julia Markovits (2010, p.204) deploys the fetishism intuition in her argument for the claim that someone performs an act with moral worth just in case she is motivated to do the morally right thing by the features that make it morally right. She too defers to Smith, arguing that someone who does the right thing because it is right "seems guilty of a kind of fetishism (to borrow a phrase from Michael Smith)" (*ibid*.). This is Markovits' main argument against the traditional Kantian idea that it might be sufficient for moral worth that an agent does the right thing because it is right.

Related ideas have spread into further literatures. For example, Alison Hills (2009, p.117, citing Arpaly 2002) appeals to the idea that agents motivated by rightness *de dicto* cannot perform morally worthy acts in a paper on moral testimony. Hills allows that someone who "wants to do what is morally right and chooses in accordance with those desires" has "good motivations", but she nonetheless insists that "more is required for morally worthy action: you need to act for the reasons why your action is right" (*ibid.*). Hills uses

this point to suggest that agents who learn what is right from testimony cannot perform acts with moral worth. Similarly, David Shoemaker (2007) argues that a "morality fetishist" – a term referring to the sort of person Smith described (*ibid.*, p.88, n.44) – would be "not responding to any moral reasons at all" (*ibid.*, p.88). This is because responding to moral reasons is, according to Shoemaker, a matter of responding to right-making features. Shoemaker takes this to show that someone motivated by rightness *de dicto* would fail to count as a full-fledged member of the moral community.

So the distinction between motivation by rightness *de dicto* and *de re*, and the associated idea that there is something wrong with motivation by rightness *de dicto*, is an old dog that is being taught new tricks. My aim in the present paper is to put a stop to this. I think that the widespread acceptance of the asymmetry thesis has been a mistake.

My own view is a form of pluralism about praiseworthy motivations. I think that it is good to be motivated by honesty, fairness, equality, and so on, *and* it is *also* good to be motivated by rightness *de dicto*. And that is not all: I hold that the traditional distinction between the right-making features and rightness itself is oversimplified. Just as there are right-making features, there are features that make it the case that the right-making features obtain – we might call them "right-making-feature-making features". And there are further features that make it the case that the right-making-feature-making features obtain, and so on, in a hierarchy of metaphysical constitution. To preview: my view is that *any* intrinsic or well-derived realizer motivation whose object is a moral feature in this metaphysical hierarchy, including the maximally thin moral feature at the top, is a praiseworthy motivation. (I will explain this all further in §2 and §4.)

Nonetheless, some of my arguments show only that certain popular criticisms of motivation by rightness *de dicto* apply with equal force to motivation by rightness *de re.*

This leaves open the possibility that neither variety of moral motivation is praiseworthy. Hence, I argue for the symmetry thesis stated conditionally: *if* motivation by rightness *de re* is praiseworthy, then so is motivation by rightness *de dicto*. My opponents already accept that motivation by rightness *de re* is praiseworthy, so I hope that they will join me in adding more praiseworthy motivations to their list. But my argument does leave open the option of throwing out both baby and bathwater and starting anew.

Here is a roadmap. I begin with three important clarifications (§2), which together suggest that there is something spurious about the distinction between two types of moral motivation on which the asymmetry thesis rests. I then argue that motivation by rightness *de dicto* and *de re* have been poorly compared, and that, when we compare correctly constructed minimal pairs, it is no longer plausible that one type of motivation is praiseworthy and the other is not. I first discuss good cases, in which people succeed in doing what they are trying to do (§3.1). I argue that the asymmetry thesis is committed to implausibly harsh verdicts about agents who try to act rightly and even partially succeed, especially as compared with those who manage to act rightly without trying. I then turn to bad cases, in which people fail to do what they are trying to do due to their false moral beliefs (§3.2). This is the main kind of case that has been used to raise worries about the praiseworthiness of motivation by rightness *de dicto*. I argue that exactly parallel worries arise for motivation by rightness *de re*. So, my opponents and I all need to find something plausible to say about such cases. I offer something to say: we must pay more attention to the different ways of being praiseworthy, acknowledging that someone can have praiseworthy motivations without praiseworthy beliefs or behavior, and that someone can have some praiseworthy motivations while lacking others. It should be no surprise that people can be criticizable in certain respects while also having some redeeming features. This, I contend, is what we should say about the well-meaning but

morally mistaken. In the final section (§4) I spell out the details of an approach to thinking about praiseworthy motivations – the "partial credit" approach – that secures this result.

2. Clarifying the phenomena

This section covers some preliminaries that are necessary for understanding my main argument. I explain the way I am thinking of motivation; I sketch the picture of moral metaphysics that informs my main argument; and I note an observation about the *de re*/*de dicto* distinction that supports the symmetry thesis, and can be overlooked by those who defend the asymmetry thesis. My disagreement with the asymmetry thesis is not only a normative one, but also one about how its alleged distinction between two types of moral motivation is set up. I think that philosophical reasoning on this point has been muddled, and that if we pay closer attention to the nature of these two types of moral motivation – to the psychology, semantics, and metaphysics – then we will see that they are have more in common than has so far been appreciated.

Here is how I am thinking of motivation. As I will construe it throughout this paper, a motivation is a type of mental state to which desire gives rise, and which itself gives rise to a set of dispositions. These comprise (1) the disposition to think about what it would take to realize that which one desires, (2) the disposition to notice when one's acts seem to have some bearing on whether that which one desires will be realized, (3) the disposition to do what one thinks will realize that which one desires, doing it *because* (one thinks that) it will realize that which one desires, and (4) the disposition to refrain from doing something if one thinks that it will impede the realization of that which one desires, refraining from doing it *because* (one thinks that) it will impede the realization of that which one desires. Someone is motivated to do something to the extent that she has these

four dispositions. For example, someone is motivated to eat healthily to the extent that she is disposed to think about healthy eating, to notice whether her food is healthy or unhealthy, and to choose to eat some foods and avoid others on the grounds that this is what it takes to eat healthily.[2]

As I am thinking of it, motivation is not quite the same thing as desire. Motivation is the part of desiring something that involves trying to bring it about. Desire itself is associated with a wider set of dispositions than the four just mentioned; for example, it is associated with the disposition to be happy and satisfied when one believes that what one desires is realized, and to be unhappy and frustrated when one believes that it is not realized. I think that it would be a conceptual stretch to say that these dispositions are part of motivation. But nothing hangs on this terminological point. If we spoke in terms of desire (or anything else) rather than motivation, it would remain the case that I am interested in the mental state that gives rise to the four dispositions just mentioned.

Some motivations are related to one another, because there are structural relationships between the desires that give rise to them. A desire to φ is *intrinsic* if it serves no further end; philosophers sometimes express this by saying that the agent wants to φ "for its own sake".[3] There are two other types of desire. A desire to φ is *instrumental* if it is generated by a desire to ψ plus a belief that φ-ing is a causal means to ψ-ing. And a desire to φ is a *realizer* desire if it is generated by a desire to ψ plus a belief that φ-ing constitutes ψ-ing. Thus, both instrumental and realizer desires depend on prior desires and beliefs about

[2] Like all dispositions, the dispositions associated with motivation need not always manifest. For example, someone could be motivated to eat healthily even though she sometimes eats cake, knowing full well that this will impede the coming about of that which she desires (viz., that she eats healthily). To the extent that she remains generally *disposed* to refrain from doing what she thinks will impede her eating healthily, and she also has dispositions (1—3), she still counts as motivated to eat healthily. These dispositions come in degrees, because motivation comes in degrees.

[3] This muddies the waters somewhat by ignoring the distinction between intrinsic and final desires, which does not matter for present purposes. For the distinction see Korsgaard (1983); Rabinowicz and Rønnow-Rasmussen (2000).

relationships between their objects and the objects of these prior desires. But they are different, since causal relationships are different from relationships of metaphysical constitution.

We can now clarify the nature of intrinsic motivation by rightness *de dicto*. This is a mental state that arises when an agent desires that she act morally rightly, and that gives rise to dispositions to think about what it takes to act rightly, to notice the moral quality of her acts, to do things she thinks are right, *because* they are right, and to refrain from doing things she thinks are wrong, *because* they would be wrong. Importantly, for a motivation to act rightly to be intrinsic, it must not depend on the agent's beliefs about what acting rightly would cause or constitute. For example, the agent is not intrinsically motivated to act rightly if she has these dispositions only because she believes that a person she finds attractive will go on a date with her if she acts rightly.

We can similarly see what it is to be intrinsically motivated by a right-making feature. For example, suppose that fairness is a right-making feature. To be intrinsically motivated by this feature is to be in a mental state that arises when the agent desires that she act fairly, and that gives rise to dispositions to think about what it takes to act fairly, to notice the fairness or unfairness of her acts, to do the things she thinks are fair, *because* they are fair, and to refrain from doing the things she thinks are unfair, *because* they are unfair. For an agent's motivation to act fairly to be intrinsic, it must not depend on her prior beliefs about what acting fairly would cause or constitute. For example, she is not intrinsically motivated to act fairly if she has these dispositions only because she believes that she will be financially rewarded for her fairness and wants some financial reward. And she is not *intrinsically* motivated to act fairly if she has the dispositions only because she believes that acting fairly constitutes acting rightly, and she wants to act rightly. (This last point

will be crucial for my argument in §3.1.) The same applies, *mutatis mutandis*, to all other right-making features.

That was the first preliminary. The second is a brief sketch of the metaphysical picture that informs my view. Most of the details of this picture are unimportant for present purposes, and could be filled out in many ways. What is important is this: *the right-making features are not fundamental*. This means that the very same metaphysical relationship – the "makes it the case" relationship – that moral rightness bears to the right-making features is in turn borne by the right-making features to various other features of acts. For example, the fact that an act is fair is not a brute fact. This fact obtains in virtue of further facts about the act; perhaps the fact that it distributes social benefits and burdens on reasonable, non-arbitrary grounds. And that is also not a brute fact. It obtains in virtue of further facts about the act; perhaps that it is meritocratic, or that it makes reparations for past injustice, or that it distributes resources based on need. And those facts are not brute, either; they, too, obtain in virtue of further facts about the act. And so on. This yields a metaphysical hierarchy of features of acts, with rightness at the top, then the right-making features, then the right-making-feature-making features, and so on down to the fundamental level. Once we zoom out and begin to see the entire hierarchy, I think that it begins to seem implausible that there is anything special about being motivated by the features at level two.

The fact that the right-making features are not fundamental means that there are *de re*/*de dicto* distinctions to be drawn with respect to motivation by any of these features, just as there is for motivation by rightness.

This is an important point about how to apply the *de re*/*de dicto* distinction. Speaking loosely, philosophers sometimes talk as if attitudes themselves carve into the *de re* and

the *de dicto*. But that is not how the distinction works. There is no such thing as a *de re* motivation or a *de dicto* motivation. Rather, the distinction applies to our *ascriptions* of attitudes.[4] *De dicto* attitude-ascriptions are given in the terms that figure in the content of the attitude. *De re* attitude-ascriptions are given in terms of things that constitute or are constituted by that which figures in the content of the attitude, where the metaphysical facts about what constitutes what are known to the person ascribing the attitude – and to her audience, if she has one – but not necessarily to the agent whose attitudes are being ascribed. For instance, a motivation to buy some flowers for Clark Kent is equally well described as a motivation to buy flowers for Clark Kent *de dicto* or as a motivation to buy flowers for Superman *de re*. A motivation to drink some H_2O is a motivation to drink some water *de re*, and also a motivation to drink some H_2O *de dicto*. And so on. As these examples illustrate, a motivation itself is neither *de re* nor *de dicto*: it is just a motivation with some content and an object, in light of which we can describe it in various ways.

This applies to motivations whose objects are features in the metaphysical hierarchy that I have described. Take any feature in the hierarchy, and take a motivation with this feature as its object. We can describe this motivation equally well in either of two ways. First, we can name the feature that is the object of the motivation, and can say that the relevant agent is motivated by this feature *de dicto*. Second, we can name the feature above this one in the metaphysical hierarchy, and say that the agent is motivated by that feature *de re*. For example, suppose that someone has an explicit concern with acting fairly: a concern with doing the fair thing in a certain situation, whatever it may be. Assuming that fairness is a right-making feature, this is one way of being motivated by rightness *de*

[4] For some background on the *de dicto*/*de re* distinction see McKay and Nelson (2014); for the most well-known analysis see Quine (1956); for an extensive contemporary treatment see Keshet and Schwarz (*ms*). Quine understands the *de re*/*de dicto* distinction in terms of the logical form of sentences, including sentences involving attitude-ascriptions. And on Keshet and Schwarz's account, noun phrases must always be interpreted as *de re* or *de dicto* relative to an intensional operator, as they may take different interpretations within the same sentence. This reinforces my point that the *de re*/*de dicto* distinction pertains to the language we use to describe attitudes, rather than to the attitudes themselves.

re. But this same motivation is equally as well described as motivation by fairness *de dicto*. By contrast, someone could care about whatever it is that fairness in fact consists in – i.e., whatever falls below fairness in the metaphysical hierarchy. This would be motivation by fairness *de re*. The same holds for all other right-making features; one can care about them *de dicto*, or *de re*, or both.

This matters because it means that motivation by rightness *de re* just *is* motivation by a right-making feature *de dicto.* To repeat: assuming that fairness is a right-making feature, being motivated by fairness *de dicto* is one way of being motivated by rightness *de re*. An explicit concern for acting fairly is equally accurately described either as motivation by fairness *de dicto* or as motivation by rightness *de re*, because these are two ways to refer to the same motivation. The same holds for all other putative right-making features. Being motivated by rightness *de re* might be a matter of being motivated to treat people with respect *de dicto*, or being motivated to promote well-being *de dicto*, or being motivated by people getting what they deserve *de dicto*. (Which of these it is depends on which moral theory is true.) And so on, for whichever features rightness in fact consists in – whatever these features are, being explicitly motivated by them is both motivation by rightness *de re* and motivation by the relevant feature *de dicto*.

What about direct concern for a loved one, rather than for an abstract idea like fairness or well-being? The same thing holds. Consider a version of the case from Charles Fried (1970) made famous by Bernard Williams (1981), in which you see that your wife is drowning and jump into the water to save her. Why do you do this? Williams famously says that your motivating thought, "fully spelled out", might have been "that it was [your] wife" (*ibid*., p.18). But this must be elliptical. Someone who came upon the scene and simply thought "Hey, there's my wife!" would not yet think anything that motivates performing any specific action. Rather, when the agent's motivating thought is *fully*

spelled out, it must be something like, "My wife is drowning! I must save her! OK, here I go!" But notice that this makes sense only if we ascribe to the agent a motivational state regarding her wife – perhaps an intrinsic motivation *to save her wife*, or perhaps a more general motivation *to protect her wife from harm* or *to care for her wife* or *to look after her wife*, from which the motivation to save her wife derives. And these are all motivations that we can ascribe to the agent *de dicto*, as I just did when listing them. These motivations are directed toward features that fall somewhere below moral rightness in the metaphysical hierarchy. But that does not prevent them from being motivations that we can ascribe to the agent in *de dicto* terms, by spelling out their content explicitly.[5]

In short, the fact that the agent cares about her wife does not make her motivation any more *de re* than any other motivation. On the contrary, it is a mistake even to think that some motivations are "more *de re*" than others. Rather, all motivations have objects and contents, in light of which we can ascribe them to agents in either *de re* or *de dicto* terms.

With these preliminaries in mind, we can now clarify the view that I oppose in this paper. This is it:

> ASYMMETRY THESIS: Intrinsic motivation by one of the right-making features *de dicto* (i.e., by rightness *de re*) is praiseworthy. But intrinsic motivation by rightness *de dicto* is not praiseworthy.

[5] To forestall a possible misunderstanding: the same thing holds if the agent's motivational state does not include the concept WIFE as part of its content, but rather a name for the wife herself – e.g., if the agent is motivated *to protect Nyika* or *to care for Ella*. These are all still motivations that we can ascribe to the agent *de dicto*, as I just did. We should not be led astray by the fact that it doesn't make sense to append "whoever that is" to these sentences, as it does for some *de dicto* expressions; e.g. we can say "amnesiac James Bond wants to protect the queen, whoever that is", but not "the agent wants to protect Nyika, whoever that is". But we can still get across the content of the agent's attitude by saying "she wants to protect Nyika, *whatever that amounts to*". This clarifies that she wants to protect Nyika *de dicto*, i.e. that she is explicitly concerned with protecting Nyika, rather than wanting to do something (e.g. shoot a pistol into a nearby alley) that happens to constitute protecting Nyika.

I think this is a faithful interpretation of what defenders of the asymmetry thesis have in mind. Defenders of this view typically explicitly restrict their focus to intrinsic motivations. For instance, in the introduction to their book, Arpaly and Schroeder say that "in this work the focus will be on intrinsic desires", and that they hold that instrumental and realizer desires have "little or [no] moral significance" (2013, p.6). There is a rationale for this restriction, which I will discuss (and criticize) in §3.1.

Defenders of the asymmetry thesis are also fairly explicit about the fact that it is the right-making features, rather than the right-making-feature-making features (or any other lower-order features), that they take to be the objects of praiseworthy motivations. When Arpaly and Schroeder say that the object of a virtuous agent's motivation is the right or good "correctly conceptualized", and that this amounts to the object of the motivation being "conceptualized in the way preferred by the correct normative theory", all of their examples are mid-level moral properties that one may care about either *de dicto* or *de re* – including "respecting persons", "happiness maximized", "welfare", and "justice" (2013, p.164). Smith's examples are also mid-level properties, like "honesty", "equality", and "people getting what they deserve" (*op. cit.*). And Arpaly and Schroeder make it clear that motivation by right-making features *de re* is not praiseworthy, on their view. They consider the case of an alien scientist who is motivated to produce high levels of activity in the perigenual anterior cingulate cortex of healthy humans, which is, in fact, what pleasure consists in. This alien is motivated to produce pleasure *de re*. But Arpaly and Schroeder say that "one would not want to credit the alien with even partial good will" (2013, p.167), even if pleasure-production is a right-making feature. So, the asymmetry thesis favors intrinsic motivation by the right-making features *de dicto*, not *de re*: this view holds that it is intrinsic motivations whose objects are right-making features, rather than right-making-feature-making features (or lower-order features), that is praiseworthy.

3. Main argument

As a reminder, here's my thesis again:

> SYMMETRY THESIS: If motivation by rightness *de re* is praiseworthy, then so is motivation by rightness *de dicto*.

I will now give my main argument for this thesis.

To assess this thesis, we should compare pairs of cases: one in which the agent is motivated by rightness *de dicto* and another in which she is motivated by rightness *de re*. But, in constructing these cases, we should tread carefully. We should ensure that we compare minimal pairs – pairs of cases in which one agent is motivated by rightness *de dicto* and the other motivated by rightness *de re*, with all else held fixed. We should avoid varying other potentially relevant factors, so as not to create noise. Notably, we should not compare one agent who tries to act rightly but has *false* beliefs about what rightness consists in, and thus ends up acting *wrongly*, to another agent who tries to perform acts with a certain right-making feature and has *true* beliefs about what it consists in, and thus ends up acting *rightly*. This is unhelpful, because our judgment about the cases does not necessarily reflect our intuitive assessment of the relative praiseworthiness of motivation by rightness *de dicto* and *de re*. It could instead be a response to other differences between the cases: perhaps the fact that one agent succeeds in what she is trying to do while the other fails, or the fact that one agent has true beliefs about the object of her motivation while the other has false beliefs about the object of her motivation, or (most probably) the fact that one agent acts wrongly while the other acts rightly.

3.1. Good cases

For a genuine minimal pair, both agents – the one motivated by rightness *de dicto* and the one motivated by rightness *de re* – should succeed in doing what they are motivated to do. They should also perform the same act under the same circumstances. I will offer one example of such a pair, and then a recipe for how to construct further examples.

Here is the example:

> CHAIRING 1: Maryam is chairing a session at a prestigious Philosophy conference, which is notorious for getting nasty during Q&A. Maryam wants to act rightly – that is, she wants to conduct Q&A in such a manner as to meet all her obligations not only *qua* chair but also *qua* moral agent. So she thinks carefully about what her obligations might be, planning to modify her behavior in light of her conclusions. After much soul-searching and careful thought, Maryam decides that four things matter morally in her case: prioritizing junior scholars over senior scholars, prioritizing those who have asked fewer questions at the conference over those who have asked lots already, discouraging audience members from asking repeated versions of the same question, and discouraging them from battering the speaker with multiple lengthy follow-ups. Maryam devises a set of principles that allows her to promote these four ends in a manner that reflects her estimation of their relative importance. She then conducts Q&A in perfect accordance with her principles. Moreover, Maryam is completely right about all of this. She has exhaustively identified the considerations that matter morally in her case, and has chosen principles that precisely

reflect their relative importance. Maryam has perfected the principles of conference ethics. Since she guides her behavior in accordance with her conclusions, she also acts perfectly.

CHAIRING 2: Mario is chairing a session at a prestigious Philosophy conference, which is notorious for getting nasty during Q&A. Mario introspects and finds that he has four intrinsic motivations relevant to his circumstances: to prioritize junior scholars over senior scholars, to prioritize those who have asked fewer questions over those who have asked lots already, to discourage audience members from asking repeated versions of the same question, and to discourage them from battering the speaker with multiple lengthy follow-ups. So Mario devises a set of principles that allows him to promote these four ends in a manner that reflects the relative degrees to which he cares about each of them. Mario also comes to believe that the objects of his motivations are the right-making features in his situation, and that it is morally right to conduct Q&A in accord with his principles, since these beliefs fit well with his pre-theoretical intuitions. But these beliefs are motivationally otiose. Mario conducts Q&A in perfect accord with his principles just because his intrinsic motivations incline him in this direction. He could change his beliefs about how it is morally right to conduct Q&A without his behavior changing at all. Happily, though, Mario's intrinsic motivations are directed toward all and only the things in his situation that are in fact morally significant, and their relative strength corresponds precisely to these things' relative importance. So, since these motivations guide his behavior, Mario also acts perfectly.

Let's assume that CHAIRING 1 and CHAIRING 2 are part of a broader pattern, as follows. Maryam has one intrinsic motivation operative in her decisions: the motivation to act rightly.[6] Mario, on the other hand, has a hodge-podge of various intrinsic motivations. But Maryam has all and only the true moral beliefs, and all and only true beliefs about morally relevant non-moral matters. So, she has realizer motivations directed toward all the right-making features, all the right-making-feature-making features, and so on. Meanwhile, Mario's intrinsic motivations happen to be directed toward all and only the right-making features. He has true beliefs about what each of these features consists in and has developed the appropriate realizer motivations. In short, for every intrinsic motivation of Mario's, Maryam has a realizer motivation with the same object. And for every realizer motivation of Maryam's, Mario has either the same motivation or an intrinsic motivation with the same object. These agents' motivational sets are almost identical. The only difference between them lies in the structure of the very top of their motivational sets: Maryam has an extra intrinsic motivation, *to act rightly*, from which her other motivations derive, while Mario's motivations derive from his intrinsic motivations directed toward the right-making features (which are, for Maryam, the objects of realizer motivations). But this difference in the structure of their motivational sets makes no difference to their behavior. In all actual circumstances, like CHAIRING 1 and 2, Maryam and Mario act identically – and, by stipulation, morally perfectly.

These cases compare two *successful* agents, who do what they are trying to do. Maryam tries to act rightly, and does a great job. She acts impeccably. Mario tries to promote each of the various things that he cares about, and does an equally great job. He promotes

[6] This is not to say that the motivation to act rightly is Maryam's only intrinsic motivation. She may have any number of other intrinsic motivations, so long as they are not operative in her decisions about what to do in the cases that make up this broad pattern. For instance, it could be that Maryam is intrinsically motivated to take care of various friends and family members, but these motivations play no part in a rationalizing explanation of her choice of chairing policy, since Maryam knows that nothing she does at the conference will affect those friends and family members.

these things to the degree to which he cares about each of them. Moreover, since Mario's motivations align with the content of the true moral theory, he too acts impeccably. So, this pair of cases is well-constructed; it compares someone motivated by rightness *de dicto* with someone motivated by rightness *de re*, holding all else fixed.

What, then, should we say about the praiseworthiness of Maryam and Mario's motivations?

The asymmetry thesis entails that Maryam's motivations are *not at all praiseworthy*. This is because of two claims that are key components of this view. First, as we have seen (in §2), the view concerns intrinsic desires. On this view, then, instrumental and realizer motivations are not the sort of thing that can be praiseworthy. Second, the view says that not just any old intrinsic motivation is praiseworthy: it says that all and only intrinsic motivations whose objects are right-making features are praiseworthy. This entails that Maryam's motivations are not at all praiseworthy. For, although Maryam is motivated by every right-making feature, those motivations are not *intrinsic*. They are realizer motivations, deriving from her intrinsic motivation to act rightly and her true beliefs that these features are what moral rightness consists in. Maryam *is* intrinsically motivated to act morally rightly. But rightness itself is not a right-*making* feature; that would be circular. (To put this another way: the "makes it the case" relation is irreflexive.) So, Maryam has no motivation that is both intrinsic and directed toward a right-making feature. Thus, according to the asymmetry thesis, she has no praiseworthy motivations.

This is not at all plausible. Maryam is a morally impeccable person. Her life consists in the performance of one morally right act after another. She also has all and only true moral beliefs. And neither her consistently right actions nor her perfectly accurate moral beliefs are a fluke; Maryam is this way because she is motivated to act rightly, which has

led to a great deal of careful thought, sophisticated reasoning, and concerted moral effort on her part. She is this way because her life is guided by an unfailing, and successful, commitment to doing what is morally required of her. There are some unappealing things about someone as morally outstanding as Maryam – one may not want to have her as one's best friend, for instance. But it is simply incredible to say that her motivations are not praiseworthy to *any* degree whatsoever.[7]

This verdict on Maryam is even less plausible when we compare it to the asymmetry thesis's verdict on Mario. On this view, though Maryam's motivations are *not at all praiseworthy*, Mario's are *fully praiseworthy*. This is because (like Maryam) he has a motivation directed toward each and every right-making feature, and (unlike Maryam) these motivations are all intrinsic. But such wildly divergent verdicts are clearly the wrong result. Maryam and Mario's motivational sets are almost identical, differing only in their structure at the very top – he with intrinsic motivations directed toward right-making features, she with realizer motivations directed toward these features, derived from an intrinsic motivation to act rightly. By stipulation, that this is the *only* difference between them. Their other realizer motivations are identical. Both act perfectly. And both have all and only true moral beliefs. If Maryam and Mario were to observe each other's behavior, or to discuss any moral issue, they may be unable to identify any difference between them. Once these cases have been constructed so as to remove other grounds for differences in praiseworthiness, this difference in the structure of the very top of Maryam and Mario's motivational sets seems far too flimsy a distinction to ground the difference between full praiseworthiness and none at all.

[7] Some philosophers have suggested that an agent like Maryam qualifies as a moral saint: see Carbonell (2013).

We can contrast motivation by rightness *de dicto* and *de re* without imagining agents as unusually morally successful as Maryam and Mario. Imagine Shmaryam and Shmario, who excel in conference-chairing but make lots of other moral mistakes. There are countless possible Shmaryams who try to act rightly but do not succeed as well as Maryam, as their moral beliefs get only part-way toward the truth, so some, but not all, of their realizer motivations are directed toward genuine right-making features. And for each Shmaryam, there is a corresponding Shmario who has *intrinsic* motivations directed toward the features that are the objects of Shmaryam's realizer motivations. (For example: suppose that there are seven right-making features, and that Shmaryam has identified two of them and developed the appropriate realizer motivations. Then Shmario is intrinsically motivated by these two right-making features, but not the other five.) We can stipulate that the agents in each pair have identical beliefs about the right-making-feature-making features, and have developed all the appropriate realizer motivations. So these agents will again perform all the same actions – some of them right, some wrong. Using such pairs of cases, we can compare motivation by rightness *de dicto* and *de re*, holding all else fixed. And for each such pair, the asymmetry thesis entails that Shmario's motivations are *somewhat* praiseworthy, but Shmaryam's are *not at all* praiseworthy. These pairs of verdicts are all implausible, given how similar the agents are, and also given that Shmaryam just does not seem like someone whose motivations are not at all praiseworthy. So this gives us a method for generating well-constructed minimal pairs, each of which provides support for the symmetry thesis: if motivation by rightness *de re* is praiseworthy, then so is motivation by rightness *de dicto*.

There is a way to modify the asymmetry thesis to avoid these implausible verdicts without conceding that motivation by rightness *de dicto* is praiseworthy. We can say that there are two types of praiseworthy motivation: intrinsic motivations whose objects are right-making features, and *realizer* motivations whose objects are right-making features.

For every intrinsic motivation directed toward a right-making feature that (Sh)Mario has, (Sh)Maryam has a realizer motivation directed toward the same feature. So the modified asymmetry thesis entails that the motivational sets of the agents in each (Sh)Maryam-(Sh)Mario pair are equally praiseworthy. The modified thesis thereby avoids the unwelcome verdicts.

But this modification is more trouble than it's worth. Arpaly and Schroeder restrict their focus to intrinsic motivations for a reason: agents can develop realizer motivations directed toward right-making features in a way that does not seem praiseworthy at all.

Here is an example:

> CHAIRING 3: Bleria is an avid devotee of a certain lifestyle blogger, whose acerbic wit and impeccable fashion sense she deeply admires. Many of her decisions are driven by just one intrinsic motivation: to emulate her favorite lifestyle blogger as closely as possible. Bleria is chairing a session at a prestigious Philosophy conference, which is notorious for getting nasty during Q&A. Fortunately, Bleria thinks she has figured out what her lifestyle blogger would do under such circumstances: Bleria thinks that the blogger would conduct Q&A in a manner that prioritizes junior scholars over senior scholars, prioritizes those who have asked fewer questions over those who have asked lots already, and discourages the audience from asking versions of the same question over and over again or from battering the speaker with multiple lengthy follow-ups. Bleria also has beliefs about the relative priority that her favorite lifestyle blogger would assign to each of these four concerns. So she devises a set of principles that she thinks embody what the blogger would do, and she acts accordingly. Moreover –

> unbeknownst to Bleria, and as a matter of indifference to her – the values that Bleria ascribes to her favorite lifestyle blogger in this case correspond perfectly to the content of the true moral theory. So, like Maryam and Mario, Bleria acts perfectly.

Again, we can imagine that this is part of a broader pattern. Across a range of cases, for every right-making feature that is the object of Mario's intrinsic motivation and Maryam's realizer motivation, Bleria has a realizer motivation directed toward this feature, derived from her intrinsic motivation to act in a way that emulates her favorite lifestyle blogger and her belief that the blogger would perform acts with this feature. But the fact that these features are *right-making* is a matter of indifference to Bleria. It is not because of the lifestyle blogger's moral character that Bleria seeks to emulate her, but because of the aesthetic appeal of her Instagram feed.

This spells trouble for the modified asymmetry thesis. The modified asymmetry thesis entails that Bleria's motivations are *fully praiseworthy*, just like Maryam's. Again, this is just not plausible. Bleria is a fashionista whose realizer motivations happen to align in content with the true moral theory. Since she is indifferent to morality *per se*, Bleria would not even be pleased to learn that the objects of her realizer motivations turn out to be all and only the right-making features. This would strike Bleria as nothing more than an amusing coincidence. This coincidental orientation toward morality does not seem very praiseworthy. By contrast, Maryam's realizer motivations are directed toward all the right-making features due to her conscientious and successful moral effort. That makes Maryam's motivational set far more praiseworthy than Bleria's. So, we cannot say that just any old realizer motivation that hits upon a right-making feature is praiseworthy. The provenance of these motivations matters. The modification fails.

We now have a recipe for constructing counterexamples to the asymmetry thesis. Here is the recipe: Pick one or more of your favorite right-making features. Imagine an agent who is intrinsically motivated by these features. Then imagine another agent who is intrinsically motivated to act rightly, has figured out that these features are right-making, and has developed the appropriate realizer motivations. Compare the two agents. Next, imagine a third agent with a morally neutral intrinsic motivation, a belief that performing acts with your preferred right-making features constitutes achieving the object of this motivation, and the corresponding realizer motivations to perform acts with these features. Compare all three agents. Et voilà! You have a dilemma for the asymmetry thesis. Unmodified, it yields implausible verdicts about the relative praiseworthiness of the first two agents. Modified, it yields implausible verdicts about the second and third.

So I submit that the asymmetry thesis should be rejected.

What caused this trouble was a pair of claims: that only intrinsic motivations are praiseworthy, and that intrinsic motivation by rightness *de dicto* is not praiseworthy. Abandoning the first of these claims alone does not help – it lands us on the Bleria horn of the dilemma. So, we should abandon the second claim too. The appropriate response to cases of people trying to act rightly and succeeding fantastically, like Maryam, is to accept that their motivations are indeed praiseworthy. Or, at least, they are praiseworthy *if* Mario's motivations are praiseworthy. This is exactly what my symmetry thesis says.

3.2. Bad cases

One popular argument against the praiseworthiness of motivation by rightness *de dicto* notes that people can be led by this motivation to act wrongly if they have false beliefs

about what moral rightness consists in. In these cases, the argument goes, the agents often don't look very praiseworthy.

I accept that motivation by rightness *de dicto* can lead someone to act wrongly, if she has false moral beliefs. But this is equally true of motivation by rightness *de re*. We can recognize this point once we understand that (as argued in §2) motivation by rightness *de re* just is motivation by one of the right-making features *de dicto*. With that in mind, notice that rightness is not the only moral property. Many "thicker" moral properties are plausible candidates for being right-making features.[8] So beliefs about these properties' nature and extension – for example, beliefs about what fairness, well-being, or justice consists in – are moral beliefs. And someone can be motivated by one of these features while being ignorant of its precise nature and extension, just as we can with respect to moral rightness. If someone is ignorant about what fairness consists in, then, by *trying* to act fairly, she can end up *in fact* acting unfairly. Parallel remarks apply to promoting well-being or justice. But it is wrong to act unfairly, undermine well-being, or inhibit justice. So, motivations whose objects are right-making features – the kind of motivation often called "motivation by rightness *de re*", but equally as well described as motivation by the relevant right-making feature *de dicto* – can lead someone to act wrongly, if she has false moral beliefs.

Here are three cases to illustrate this point:

> FAIRNESS: A father is coming up with a toy-sharing policy for his two daughters. He wants his toy-sharing policy to be fair. So he thinks awhile and comes up with a rudimentary theory of fairness. But he gets it wrong;

[8] "Thick" properties are those denoted by thick concepts, which are partly descriptive and partly normative. See Roberts (2013) for an introduction, and Väyrynen (2013) for detailed discussion.

he thinks that his daughters' age-difference is irrelevant to considerations of fairness, when in fact it is relevant. So he ends up instituting a policy that is in fact unfair to his younger daughter.

WELL-BEING: A mother wants to promote her son's well-being. She thinks it will promote his well-being for him to learn a musical instrument. So she signs him up for piano lessons and forces him to go. But this underestimates the importance of autonomy as a component of well-being; the son doesn't want to learn piano, so her forcing him to do it in fact undermines, rather than promoting, his well-being.

JUSTICE: Some parents are trying to think of a just punishment for their child, who has drawn on the walls of their house. They falsely believe that smacking is, sometimes, a just punishment. And they believe that this is one of those times. So they smack their child. But they are wrong; smacking is never a just punishment.

Faced with cases like these, it is tempting to say that the parents are at least praiseworthy for *trying* to create a fair toy policy, promote the son's well-being, and come up with a just punishment, even if they are also blameworthy for *in fact* acting unfairly, undermining well-being, and inhibiting justice. I think that this is the right thing to say. But if we say that about these cases, then we can say the same thing about trying and failing to act rightly. We can imagine analogues of FAIRNESS, WELL-BEING, and JUSTICE in which the agents want to act *rightly* and falsely believe that it is *right* to institute the toy policy, force the son to take piano lessons, or smack their child. Once we construct our cases in this way – as genuine minimal pairs – it is no longer plausible that trying and failing to act rightly is *ipso facto* less praiseworthy than trying and failing to act, say, fairly.

In all cases, we have people who are well-meaning but morally mistaken. It just doesn't seem to make a difference whether they are mistaken about rightness or another moral property. J.S. Mill famously remarked that "there is no difficulty in proving any moral standard whatsoever to work ill, if we suppose universal idiocy to be conjoined with it" (Mill 1871, p.35); the same holds of moral motivations.

This supports the symmetry thesis. If motivation by rightness *de re* is praiseworthy even when led astray by false moral beliefs, then so is motivation by rightness *de dicto*. And if motivation by rightness *de dicto* is no longer praiseworthy when led astray by false moral beliefs, then the same goes for motivation by rightness *de re*.

Here my opponents might object. Several authors have argued that false moral beliefs cannot excuse an agent from blame for wrongdoing, and that agents who are led to act wrongly by their false moral beliefs are still blameworthy (see e.g. Harman 2011). My opponents might worry that the position I am defending challenges this view, by suggesting that such agents might, in fact, be praiseworthy.

This worry is misplaced. The question of whether moral ignorance excuses is a question about when agents are blameworthy *for acting wrongly*. To answer it, we may need to know when agents are blameworthy *for moral ignorance*. But the symmetry thesis is about praiseworthy *motivations*. And motivations, beliefs, and acts are all different things. So our verdicts about them can come apart: someone may be blameworthy for one or two of them while being praiseworthy for the rest. This means that we can say that, in cases like FAIRNESS, WELL-BEING and JUSTICE, the agents are praiseworthy for *trying* to act fairly, promote well-being, or bring about justice, even if they are blameworthy for *in fact* acting unfairly, undermining well-being, or inhibiting justice. We can even say that someone is still praiseworthy for her good motivation if she is blameworthy not only for her

wrongful act, but also for her false moral belief. For example, it might be that the parents in JUSTICE are blameworthy both for thinking that smacking is permissible (thus displaying insufficient concern for the child's welfare) and for smacking their child, but nevertheless are praiseworthy *for wanting to find a just punishment when their child has drawn on the walls*. Even if the act and belief are blameworthy, the good motivation may still be praiseworthy.

I think that this is the right thing to say about these cases. It is natural to say, "Her intentions were good", taking oneself to be mentioning a redeeming feature of an agent who has acted poorly. I think such claims are often literally true. The agent's intentions *were* good – that is to say, her motivations were praiseworthy. It is a commonplace that we can be criticizable in some respects while also having some redeeming features; people are not either wholly perfect or wholly awful. I am suggesting that this holds of the well-meaning but morally mistaken. Good motivations can still be praiseworthy even in agents who act poorly or hold false moral beliefs.

And if this holds for motivation by rightness *de re*, then it should also hold for motivation by rightness *de dicto*. The fact that someone was at least *trying* to act rightly can be a redeeming feature just as well as the fact that she was at least *trying* to act fairly, in cases where the agent ends up acting wrongly due to false moral beliefs. So this kind of case provides further support for the symmetry thesis: once again, motivation by rightness *de re* and motivation by rightness *de dicto* seem perfectly analogous.

My opponents may now raise a different worry. Arpaly and Schroeder (2013, pp.186-7) note that false moral beliefs can erode, and eventually eliminate, someone's praiseworthy motivations. They imagine someone who is initially intrinsically motivated by a right-making feature, but who becomes convinced of a false moral theory, and is also

motivated by rightness *de dicto*. They imagine that the agent is then so concerned to act well by the lights of her false theory that the intrinsic motivation directed toward that which truly is right-making loses its grip on her. We might say that such agents are literally *corrupted by theory*.

I agree that people can lose praiseworthy motivations when they are corrupted by theory. But this concern does not raise doubts about the value of motivation by rightness *de dicto*, nor about the symmetry thesis. That is because the risk of people's being corrupted by theory is not confined to motivation by rightness *de dicto*. It arises with equal force for motivation by rightness *de re* (i.e., motivation by one of the right-making features *de dicto*). For example, the parents in JUSTICE may be led by their false theory of justice to slowly lose their natural inclinations against hitting their child. Or the father in FAIRNESS may find his inclination to be more lenient with his younger daughter slowly dissipating as he becomes increasingly convinced of his false theory of fairness. Again, we may want to say that these agents are still praiseworthy for *trying* to bring about justice or to institute a fair toy policy, even though this blinds them to their initial concern for that which justice and fairness actually consist in. And, again, it is hard to see why we should not then say the same thing about trying to act rightly. Again, then, symmetry persists.

Here is a third, related, worry. My opponents may suggest that some agents' moral beliefs are so wildly askew that they deserve *no praise whatsoever* for trying to act rightly. If someone's conception of what is morally right is way off-track, and this leads them to commit horrific acts, then perhaps it is implausible that their motivation to act rightly is still praiseworthy.

Arpaly (2003, pp.98-101, 111-114) says roughly this. She considers and rejects the possibility that it may be a virtue to "stick to one's guns" – i.e., to get oneself to do what

one believes to be morally right – if one is morally mistaken. Similarly, here is Markovits (2010, p.224):

> [T]he fact that Göbbels was driven by his conscience to persecute the Jews does not exonerate him, much less endow his acts with moral worth.

Markovits is here arguing that, if Göbbels was trying to act rightly, this should not make us think better of his wrongful act. But we can equally imagine someone arguing that there is nothing of value in Göbbels' motivations, notwithstanding the fact that he wants to act rightly and believes that what he is doing is right. Perhaps if someone gets *really* bad, then their so-called "good intentions" are no longer a redeeming feature.

Let's be clear about what we are being asked to imagine here. We are being asked to imagine someone who sincerely believes that she is morally required to perform acts that are in fact completely horrific. That is false of many historical figures who *claim* to have sincerely believed that moral atrocities they committed were morally required. People often use moral language to advance their own interests; they use positively-valenced moral terms to describe terrible acts that they perform, order, or sanction, without believing the claims that they are making, in the attempt to manipulate others and thereby to gain and maintain power. People also use positive moral language to describe terrible acts that they perform, order, or sanction so as to convince *themselves* that these acts are not so bad, reducing cognitive dissonance. So the use of positive moral language by agents perpetrating moral atrocities does not show that these agents care *de dicto* about rightness, fairness, justice, or anything else. It could instead suggest that people mask behavior that they know to be morally atrocious in positive terms in order to sleep at night.

It is also worth remembering that we are being asked to imagine someone who sincerely believes that he is morally required to perform horrific acts, yet is sufficiently competent with moral language to be able to refer to the property of moral rightness (and thus to be motivated by rightness *de dicto*). An agent may successfully refer to such properties as rightness, justice, or desert, even if she has an incomplete understanding of their natures. But when her beliefs about these properties are *really* warped, she may fail to refer to them at all. This stems from a general feature of reference. For example, suppose that someone claims to want to visit "Detroit", but that her *only* belief about Detroit is that it is somewhere in England. There may be somewhere that this person wants to visit, which she calls "Detroit". But it is not Detroit; she fails to refer to Detroit. Likewise, someone who claims to care about a thing that she calls "rightness", but whose *only* belief about rightness is that it is a property of her left shoe, fails to refer to rightness. If this line of reasoning is correct, then some agents whose moral beliefs are wildly askew may fail to be motivated by rightness *de dicto* at all. They are motivated by something, which they call "rightness". But it is not rightness.

What we are being asked to imagine, then, is an unusual sort of person. We are being asked to imagine someone who is competent with moral concepts, understanding enough about moral rightness to count as being motivated to act rightly, and we are being asked to imagine that she sincerely believes of some moral atrocities that they are morally required. She isn't faking, or engaging in self-deception, or using the word "right" to mean something else. On the contrary, she is honestly trying to do what's morally right. But she has been led – presumably either by very misleading evidence or catastrophically terrible reasoning – to believe that what is morally right is something that is in fact morally atrocious.

I am inclined to bite the bullet at this point. That is, I am inclined to say that this agent *is* praiseworthy for sincerely trying to act rightly. Notice that we can still say that this motivation is the *only* praiseworthy thing about her; we can still say that she is blameworthy for her wrongful acts and false moral beliefs. So, even if her moral motivation is praiseworthy, she may still be an utterly despicable person overall. Given that, I do not find this the hardest bullet to bite. Indeed, we can motivate the claim that sincere moral motivation is a redeeming feature even of someone like Göbbels using another minimal pair: we can compare a clueless Göbbels who sincerely believes that his actions are right with a knowing Göbbels who is fully aware that his actions are deeply wrong and just doesn't care. This is a difficult comparison. But I am tempted to think that the former agent is at least slightly better than the latter. If so, that is presumably because his intentions are good.

Perhaps you disagree. You may be persuaded that, if someone gets *really* bad, then their so-called "good intentions" are no longer a redeeming feature. In that case, I think you still can and still should accept the symmetry thesis. Recall that this thesis states that *if* motivation by rightness *de re* is praiseworthy, then so is motivation by rightness *de dicto*. Motivation by rightness *de dicto* may not always be praiseworthy, and circumstances involving wildly askew moral beliefs may be among the times when it is not. A problem for the symmetry thesis would only arise if motivation by rightness *de re* is praiseworthy *under the same circumstances* – holding everything else fixed. And I very much doubt that this is the case here. Once again, we can imagine cases of being misled by motivation by rightness *de re* and false beliefs about what the right-making features consist in. Suppose that Göbbels wanted to act *justly* and thought it *just* to persecute Jews, rather than right. Or suppose that he was motivated by the thought of *people getting what they deserve*, and thought that Jews deserve persecution. (Either supposition might accurately characterize the actual historical Göbbels.) These versions of Göbbels seem no better than the version

who is motivated by a drastically mistaken conception of moral rightness. Either way, his moral beliefs are wildly askew and his actions unconscionable, to the point where his caring about something that it is usually good to care about may not redeem him. So in this case, again, whether the agent is (mistakenly) motivated by rightness *de dicto* or *de re* simply does not matter. Symmetry persists.

4. The "Partial Credit" Approach

It is possible for defenders of the asymmetry thesis to take a hard line on cases of well-meaning but morally mistaken agents. They can say that agents are praiseworthy only if their motivations align precisely with the *true* nature and extension of the right-making features. On this approach, an agent's being praiseworthy requires more than that (e.g.) fairness is a right-making feature and she is motivated by fairness *de dicto*. On this approach, she must *also* have true beliefs about what fairness consists in, and must have developed the corresponding realizer motivations. Moreover, she must have still further true beliefs about that which that-which-fairness-consists-in *itself* consists in, and must have developed the corresponding realizer motivations. (For example, if fairness consists in distributing benefits and burdens on reasonable grounds, then she must have a realizer motivation to distribute benefits and burdens on reasonable grounds, and she must have true beliefs about what this amounts to, and she must have realizer motivations directed toward whatever it amounts to.) And so on, all the way down the metaphysical hierarchy discussed in §2. And so, similarly, for any other right-making feature. Call this "the hard-line approach".

The hard-line approach entails that those who are led to act wrongly by their false moral beliefs, including those who are corrupted by theory, do not have praiseworthy

motivations. On this view, it is simply false to say of the agents in FAIRNESS, WELL-BEING, and JUSTICE that their intentions were good. Their intentions *would* have been good if they were accompanied by true beliefs and realizer motivations about that which the objects of the intentions consist in (and about that which the features that they consist in consist in, etc.). But these agents have false beliefs, and their realizer motivations do not align with the true nature of these features. So, on this view, their motivations are not praiseworthy.

But the hard-line approach is unappealing, since it is likely to entail that no actual person has praiseworthy motivations. Whether it does so depends on what moral theory turns out to be true; the hard-line approach will grant moral praiseworthiness only if the right-making features turn out to be things that everybody is moved by and that we all understand perfectly. This is, of course, exceedingly unlikely. For any plausible candidate for being a right-making feature, normal people have only an inchoate grasp of this feature, rather than detailed beliefs about its precise nature and extension with a full set of realizer motivations. For example, many people are intrinsically motivated by justice *de dicto*. But I take it that nobody simply finds themselves naturally motivated by each person's having the highest degree of basic liberties compatible with equal liberty being granted to all and by social and economic inequalities' being (a) distributed to benefit the least well-off and (b) open to all under conditions of fair equality of opportunity, with (b) taking lexical priority over (a). Yet the most famous and influential theory of justice (Rawls 1971) says that this is what justice consists in. If anything like this theory is true, then, it follows that only a few people – who read Rawls, were persuaded, and remember his account in detail – are at all motivated by justice *de re*. This generalizes: the true moral theory, fully spelled out, would provide us with accounts of the nature and extension of the right-making features that far surpass ordinary agents' understanding of them and are objects of motivation for nobody. So, by making this understanding and these

motivations a necessary condition of our ordinary motivations' being praiseworthy, the hard-line approach effectively rules that our ordinary motivations are not praiseworthy.

People's motivations are often praiseworthy. So we should not take the hard-line approach.

Instead, I propose that we take what I will call a "partial credit" approach. This approach says that we are praiseworthy for having motivations whose objects *approximate* the content of the true moral theory, and we are more praiseworthy the closer the approximation is.

Here is what that means. In §2 I described a metaphysical hierarchy of right-making features, right-making-feature-making features, and so on. The true moral theory, fully spelled out, would tell us a large part of what this hierarchy is. It would exhaustively specify the right-making features, clarifying the relationships between them and any conditions on their being right-making, and it would tell us what metaphysically constitutes these features. At least, it would tell us these things about those of the right-making features that are themselves moral features – honesty, equality, desert, fairness, well-being-promotion, and so on. These being moral features, the task of specifying their nature and figuring out what it takes for them to obtain is part of moral theory. So, when I say that people are praiseworthy for having motivations whose objects approximate the content of the true moral theory, I mean that people are praiseworthy for having motivations whose objects are the moral properties in this metaphysical hierarchy. And when I say that people are more praiseworthy the closer the approximation is, I mean that someone is more praiseworthy, the more of the moral properties in this hierarchy are objects of motivation for her.

For example, continue to suppose that fairness is a right-making feature. Then the partial credit approach says that anyone motivated to act fairly *de dicto* is somewhat praiseworthy. But she is *more* praiseworthy the more accurate her conception of fairness is, and thus the more her realizer motivations align with the true nature of fairness (i.e. are directed toward the properties falling below fairness in the true metaphysical hierarchy.) Now return to the father in FAIRNESS. He does not have realizer motivations that align with the true nature of fairness, so he is not as praiseworthy as he could be. But he is, at least, *trying* to act fairly. So he gets partial credit; his intrinsic motivation directed toward fairness *de dicto* is a praiseworthy motivation. Assuming that fairness is a matter of distributing benefits and burdens on reasonable grounds, the father gets a bit more credit; he has figured this much out, and has developed a realizer motivation to distribute benefits and burdens – in this case, toy playtime – on reasonable grounds. That is a further praiseworthy motivation, on the partial credit approach. This is so despite the fact that the father is mistaken about what sorts of grounds are reasonable, such that his realizer motivations from this point on diverge in content from the true moral theory and thus are not praiseworthy.

Here is another example. Imagine someone who cares about fairness, knows that it consists in distributing benefits and burdens on reasonable grounds, and has developed a realizer motivation to distribute benefits and burdens on reasonable grounds. But suppose our agent thinks that considerations of increased future utility are the only reasonable grounds on which to distribute benefits and burdens. Let's stipulate that she is wrong about this: in fact, considerations of increased future utility are *among* the reasonable grounds, but are not the whole story – there are also considerations of merit, and of reparations for past injustice. Our agent then gets partial credit. She is motivated to act fairly, to distribute benefits and burdens on reasonable grounds, and to take considerations of increased future utility into account. All of this is praiseworthy, on my

view (given our moral stipulations). But the agent would be *more* praiseworthy were she *also* motivated (a) to take merit into account and (b) to make reparations for past injustice. That is how the partial credit view works.

We need an important qualification at this point. Not just *any* motivation whose object is one of the moral features in the hierarchy is a praiseworthy motivation. Our motivations' provenance matters; this was one of the lessons of §3.1. A motivation with one of the moral features as its object is praiseworthy if it is *either* an intrinsic motivation *or* a realizer motivation deriving from an intrinsic motivation directed toward a moral feature further up in the hierarchy, plus true beliefs about the metaphysical relationships that hold the hierarchy together. (Call the latter a "well-derived" realizer motivation.) For example, take someone motivated to distribute benefits and burdens on reasonable, non-arbitrary grounds. She is praiseworthy for this if she cares about it intrinsically, or if she cares about it because she cares about fairness and knows that this is what fairness consists in, or if she cares about it because she cares about acting rightly, knows that fairness is a right-making feature, and knows that this is what fairness consists in. But the agent is not praiseworthy for caring about distributing benefits and burdens on reasonable, non-arbitrary grounds if she does so because she is intrinsically motivated to imitate her favorite lifestyle blogger or to secure the attraction of her beloved, or the like, and she believes that distributing benefits and burdens on reasonable grounds constitutes attaining one of these non-moral goals. Realizer motivations whose objects are moral features in the hierarchy are praiseworthy when they are based on accurate – though perhaps incomplete – appreciation of the moral significance of their objects.

A stronger version of this qualification applies to motivation by non-moral features that may appear lower down in the metaphysical hierarchy.[9] For these non-moral features, I suggest that only well-derived realizer motivations are praiseworthy. It would be odd, and not especially praiseworthy, for these features to be the objects of *intrinsic* motivation. For example, it is praiseworthy to have a realizer motivation to ensure that we all have plenty of oxygen to breathe, having recognized that this is a vital human need and being intrinsically motivated to contribute to the satisfaction of people's needs. But it is odd, and not particularly praiseworthy, to be *intrinsically* motivated to make sure people have plenty of oxygen to breathe. Divorced from any beliefs about the value of oxygen for humans, wanting to ensure that people breathe plenty of oxygen *for its own sake* would just be weird. And this generalizes. So, it is praiseworthy to have well-derived realizer motivations directed toward non-moral features that realize the moral features in the hierarchy, but it is not praiseworthy to have intrinsic motivations directed toward the non-moral features. In general, an agent is praiseworthy for caring about something that matters morally iff she has figured out at least part of the story about *why* it matters morally, and she cares about it on this basis.

These remarks invite a final important class of objections. The reader may think that my remarks about motivation by the non-moral features in the hierarchy apply equally to motivation by rightness *de dicto*. I have suggested that it is not praiseworthy to care about one of the non-moral features with no appreciation of the moral features above it in the metaphysical hierarchy, as these are what make it morally significant. The reader may think that, similarly, it is not praiseworthy to care about rightness with no appreciation

[9] Whether there are non-moral features in the hierarchy, and how low-down they are, depends on whether moral naturalism is true. For example, if the sort of robust realism defended by Enoch (2011) is true, then the moral is fundamental, and there is no level in the hierarchy such that the levels below it contain only non-moral features. The same holds if the right-making features include thick properties (as I have assumed) and the "anti-disentanglement" argument about such properties is correct; on this see Roberts (2013), pp.680-681; McDowell (1998); Putnam (2002). By contrast, if moral naturalism is true, then non-moral features will show up in the hierarchy at some point.

of the moral features *below* it in the metaphysical hierarchy – the right-making features, right-making-feature-making features, etc. – as these are the features that lend moral rightness its significance. She may say that what makes moral rightness significant are the features that it consists in, and that one fails to see why moral rightness matters if one does not appreciate these features. So, she may say, intrinsic motivation by rightness *de dicto* is not enough for praiseworthiness: it must be accompanied by appreciation of the right-making features, otherwise praiseworthiness would be too easy to attain. Relatedly, the reader may worry that motivation by rightness *de dicto* with no appreciation of the right-making features would be empty of content, and unable to guide action. Or she may wonder how someone could come to have this motivation.

The first thing to note about these objections is that – as usual – they generalize. Parallel worries arise for intrinsic motivation by any of the right-making features *de dicto*, just as for motivation by rightness *de dicto*. For example, take kindness. I am unable to articulate exactly what kindness consists in. But I do care about treating others with kindness, and I want to act kindly. I assume that I am not alone in either of these respects. The reader may now allege that it is not praiseworthy to care about kindness with no appreciation of the features below it in the metaphysical hierarchy – of kind-making features, kind-making-feature-making features, etc. She may say that it is the things kindness consists in that make kindness morally significant, and that we fail to see why kindness matters if we do not appreciate these features. So, she may say, intrinsic motivation by kindness *de dicto* is not enough for praiseworthiness: it must be accompanied by appreciation of the kind-making features, otherwise praiseworthiness will be too easy to attain. Relatedly, she may worry that motivation by kindness *de dicto* with no appreciation of kind-making features would be empty of content, and unable to guide action. Or she may wonder how someone could develop this motivation.

I think that both sets of objections are wholly mistaken. I will now explain why.

Contrary to what the objections suggest, it is not all that easy to have one of the moral properties in the true metaphysical hierarchy as the object of one's intrinsic motivation. First, for anything to be the object of an agent's attitude, the agent must be able to refer to it (as we saw in §3.2). This places constraints on what counts as an intrinsic motivation directed toward a feature in the hierarchy *de dicto*. If someone says that she cares about acting X-ly, but can say nothing whatsoever about X, then there is nothing to make it the case that her term "X" refers to rightness, kindness, or anything else. So, the agent must grasp *something* about the object of her motivation if she is to refer to it, and thus if it is to be the object of her motivation.

But the agent need not attain this grasp by knowing what falls below the feature in the true metaphysical hierarchy. There is a difference between the nature of moral rightness and the things that rightness consists in. This explains why people with starkly different beliefs about what the right-making features are can substantively disagree with one another, rather than talking past each other.[10] These people disagree about what the property of moral rightness consists in, but they share an understanding of its nature. The same holds for disagreement over what constitutes one of the right-making features. Substantive disagreement is possible because people can share an understanding of the nature of these features, thereby talking about the same thing, while disagreeing about what it consists in.

There are ways of elucidating the nature of moral rightness that remain neutral as to what the right-making features are. For example, someone might characterize the morally right

[10] As is the standard intuition about "moral twin Earth" cases; see Horgan and Timmons (1990, 1992).

as that which would secure the agreement of all reasonable persons, or that which a suitably idealized observer would recommend, or that which we would be subject to fitting blame for failing to perform. Or she might characterize moral rightness as the property of being required by the true moral theory, of being supported by the balance of moral reasons, or of responding adequately to all the morally significant features of a situation – provided that she understands enough about the nature of morality to distinguish *moral* theory, reasons, and significance from other kinds of theory, reasons and significance (of prudence, say, or of nutrition). For present purposes we need not settle the question of whether any of these glosses on the concept of moral rightness is correct,[11] nor the questions of how many and which glosses someone must have in mind in order to grasp the concept of moral rightness. I think it is plausible that the concept MORAL RIGHTNESS is a cluster concept. But someone must grasp *something* along these lines for moral rightness to be the object of her motivation. So, her motivation cannot be completely empty of content. Once again, parallel remarks apply to each of the right-making features.

There is another sense in which motivation by one of the moral features in the true metaphysical hierarchy is not easy to come by. Motivation by rightness *de dicto*, for example, requires a lot more than sitting around saying "I love rightness". Motivation is a complex mental state giving rise to the four dispositions discussed in §2. Someone is motivated by rightness *de dicto* to the extent that she is disposed to think about what rightness consists in, notice the moral quality of her acts, and choose to perform some acts and refrain from performing others on the grounds that doing so is morally right. Similarly, an agent is motivated by fairness, kindness, well-being, justice, or any other right-making feature to the extent that she is disposed to think about what it consists in,

[11] For defenses of various versions of these glosses, see Railton (1989), (1993); Gibbard (1990); Smith (1994); Darwall (2010); Scanlon (1998); Stratton-Lake (2002); Cuneo and Shafer-Landau (2014).

to notice which of her acts possess it, and to choose to perform some acts and refrain from performing others on this basis. These dispositions come in degrees, because motivation comes in degrees. And a motivation's praiseworthiness can also come in degrees, corresponding to its strength. A weak motivation is only weakly praiseworthy. And it takes quite a lot for someone to count as strongly motivated by many of the features in the true metaphysical hierarchy; she would have to display the four dispositions above to a high degree with respect to many of these features. This means that it takes more to be *highly* praiseworthy, on the partial credit approach, than a cynical reader may imagine.

These points help us to see how someone can become intrinsically motivated by a feature in the hierarchy. Return to the case of fairness. Someone may initially be intrinsically motivated by meritocracy, reparations for past injustice, and distributing resources based on need. After reflecting on these features and on what makes them morally significant, she may come to grasp that there is something they have in common: they are all ways of distributing benefits and burdens on reasonable, non-arbitrary grounds, which is what *fairness* is. She can thus re-conceptualize the objects of her prior motivations, seeing them all as realizers of fairness, where before she saw them as independently intrinsically valuable. She can then kick away the ladder: that is to say, she can decide that it is *acting fairly* that she really cares about, whether or not it turns out to consist in precisely the three things with respect to which she initially grasped this feature. Similarly, someone may be initially intrinsically motivated by fairness, kindness, honesty, and the like, and after some reflection may re-conceptualize them all as realizers of a larger thing: *moral rightness*. She can then decide that it is *acting rightly* that she really cares about, whether or not it turns out to consist in precisely the things with respect to which she initially grasped this feature. Thus, agents can develop motivations whose objects are features in the metaphysical hierarchy (e.g. fairness or rightness) by abstracting away from concerns for the more concrete things that these features consist in (e.g. meritocracy and

reparations, or kindness and honesty), re-conceptualizing the objects of their concerns as realizers of a broader moral category. Once the agent starts to think that she cares about *whatever* falls within this broader moral category, whether or not it is exhausted by the lower-order features that she initially had in mind, she has become intrinsically motivated by the relevant property *de dicto*. So, it is no mystery how agents can develop intrinsic motivations with moral properties as their objects. On the contrary, these reconceptualization processes form a major part of the process of moral education.

This picture of how reconceptualization processes work highlights two advantages of the partial credit approach. First, on my approach, we can have praiseworthy motivations throughout a reconceptualization process. Intrinsic and well-derived realizer motivations directed toward moral features in the hierarchy are both praiseworthy, according to the partial credit approach. This means that someone's motivation to perform acts with one such feature remains praiseworthy if it alters from being intrinsic to being a well-derived realizer motivation (for example, if she ceases to care about meritocracy intrinsically and comes to see it as valuable *qua* realizer of fairness). And the intrinsic motivation directed toward a higher-up moral feature in the hierarchy that the agent develops through this process is also praiseworthy.

The partial credit approach's lenience on this point makes it a good fit with our ordinary intuitions about who is praiseworthy. On both the partial credit and hard-line approaches – and according to all forms of the asymmetry thesis – which motivations are praiseworthy depends on which moral theory is true. But the partial credit approach is more lenient, as it allows motivations whose objects are right-making-feature-making features (or lower-order features) to be praiseworthy, and it allows both intrinsic and well-derived realizer motivations to be praiseworthy. This accords well with our actual practices of evaluating people's motivations. It is natural to think that one can tell roughly

which motivations are praiseworthy even if one does not know precisely which moral theory is true. I suggest that this is because, even without knowing precisely which moral theory is true, one can be fairly confident that certain features of acts have *some* moral significance – that is to say, one can be fairly confident that they fall *somewhere* in the true metaphysical hierarchy. This is enough, according to the partial credit approach, to be fairly confident that an intrinsic or well-derived realizer motivation with this feature as its object is among the praiseworthy motivations. So, on my approach, we need know only a little about the structure of the true moral theory, and only a little about the structure of someone's motivational set, to get a rough sense of her praiseworthiness.

Another advantage is that the partial credit approach explains how we can improve the praiseworthiness of our motivations by engaging in moral reflection. Beginning with an inchoate set of moral concerns, and then thinking carefully about the objects of one's concerns, someone can "fill out" parts of the hierarchy and thereby develop intrinsic or realizer motivations directed toward more features within it. One way this can happen is for the agent to think about the object of a motivation, figure out what it consists in, and develop well-derived realizer motivations directed toward features falling below it in the true hierarchy. I hold that these well-derived realizer motivations are praiseworthy. Another way to fill out the hierarchy is to undergo the kind of reconceptualization process just described, wherein intrinsic motivations directed toward lower-down features in the hierarchy are transformed into realizer motivations when the agent grasps the fact that their objects are all realizers of a less fundamental moral feature that she begins to care about directly. The upshot of this process is that the agent develops an intrinsic motivation whose object is one of the moral features in the hierarchy. That is a

praiseworthy motivation. Thus, the partial credit approach offers a natural picture of how processes of moral education can improve the praiseworthiness of our motivations.[12]

There is a lot of work still to be done in working out how to apply the partial credit approach to real agents. All real people clearly fall far short of full credit: our motivations are directed toward some but not all of the features in the true metaphysical hierarchy, and also come in degrees. We need a way to compare the amounts of credit that different people get when their motivations have different strengths and align with different parts of the true moral theory. Some of these motivations may count for more than others, if their objects are more important. In that case, the total praiseworthiness of someone's motivational set will be a weighted sum of the strength of each of her motivations whose object is a moral feature in the true metaphysical hierarchy, weighted by the feature's importance. Working this all out is far beyond the scope of the present paper, as it would require us to exhaustively identify the right-making features and to determine their relative importance. This amounts to completing first-order normative ethics. Moreover, the question of how to calculate total praiseworthiness arises whether or not the partial credit approach is correct, and whether or not motivation by rightness *de dicto* is praiseworthy, so long as there are at least two praiseworthy motivations. Even someone who accepts the asymmetry thesis and the hard-line approach still faces the daunting task of calculating and comparing overall praiseworthiness if there is more than one right-making feature. So I am not worried about facing this task. It is more complicated on my approach than on some others. But it is a task that everyone faces.

[12] This is not to say that, for someone seeking to improve her overall praiseworthiness, the best means is *always* to engage in moral reflection with a view to filling out more of the hierarchy and thereby developing more praiseworthy motivations. We are praiseworthy for our conduct as well as for our motivations. And, for many people, it is a surer route to moral self-improvement to work on acting rightly more of the time, combating laziness and *akrasia* – that is to say, it is a surer route to increased praiseworthiness to work on putting one's motivations into action than on developing more of them. Thanks to an anonymous reviewer for prompting me to think about this.

5. Conclusion

Motivation by rightness *de dicto* looks bad if we compare an agent who is trying to act rightly and failing with one who is trying to perform acts with a certain right-making feature and succeeding. But these cases are not minimal pairs. They vary in whether the agent is succeeding or failing at what she is trying to do, and, crucially, in whether she is acting rightly or wrongly. They do not isolate the key issue of motivation by rightness *de dicto* vs. *de re*.

I have argued that, when we compare correctly constructed minimal pairs, motivation by rightness *de dicto* looks every bit as praiseworthy as motivation by rightness *de re*. To deny this yields unduly harsh verdicts about agents who try to act rightly and even partly succeed, especially as compared with those who manage to act rightly without trying. The asymmetry thesis entails that the motivations of people like Maryam are not at all praiseworthy, while those of people like Mario are fully praiseworthy. These extreme differences in praiseworthiness seem arbitrary and unmotivated. We can avoid this by saying that realizer motivations whose objects are right-making features are praiseworthy, but this yields unduly positive verdicts about people like Bleria. We should instead hold that motivation by rightness *de dicto* is praiseworthy.

Turning to cases involving agents who try to act rightly but fail, I have argued that all reasons to question the praiseworthiness of their motivations apply equally well to agents who are motivated by right-making features but fail to perform acts with these features. Any of these motivations can lead a person with false moral beliefs to act wrongly – including deeply wrongly, if her moral beliefs are way off-track – and any of them can "corrupt" a person by eroding her instinctual concern for that which really does matter.

We should respond to these observations by distinguishing the praiseworthiness of motivations, acts, and beliefs, acknowledging that someone's good intentions can be a redeeming feature even if she believes and acts poorly, and that someone can have some praiseworthy motivations while lacking others.

Lastly, I have argued that, in evaluating agents who are motivated by some but not all of the features in the true metaphysical hierarchy, we should take a "partial credit" approach. The partial credit approach's evaluations of real people are more lenient than those of the hard-line approach, and this is more so the more people there are who care about some moral features while failing to grasp the nature and extension of these features in full. The partial credit approach gives these normal people credit for those of their motivations that do align in content with the true moral theory, while the hard-line approach does not. But most people have only an inchoate grasp of the right-making features, which leads them to make very many moral mistakes. So the hard-line approach implausibly entails that almost no-one has praiseworthy motivations, while the partial credit approach lets us recognize the extent of each agent's moral success.

This paper is for the souls whose intentions are good (*de dicto*). I hope they will no longer be misunderstood.

II: We Can Have Our Buck and Pass It, Too

1. Setting things up

In this paper I argue that the moral rightness of an action is a reason to perform it.

When I say "moral rightness" (or just "rightness"), I mean the property of being required by the true first-order moral theory. I remain neutral as to what this theory is.

When I say "reason", I mean an objective normative reason. So when I say that an action's moral rightness is a reason to perform it, I mean that the action's rightness counts in favor of performing it, regardless of whether anyone is aware that the action is right and regardless of whether its rightness motivates anyone to perform it.[13] By analogy: if an island is beautiful, this is a reason to visit it – it counts in favor of visiting it – regardless of whether anyone is aware that it is beautiful and regardless of whether its beauty motivates anyone to visit it.

As the above paragraph makes clear, in this paper I avail myself of a popular claim about objective normative reasons:

[13] For brevity, I will often speak of a property of an act (e.g., rightness) being a reason to perform it. But I am happy to make the orthodox assumption that normative reasons are facts, not properties. When I say that a property is a reason, this should be understood as elliptical for the claim that *the fact that* an act has some property is a reason to perform it.

REASONS: A reason to φ is a consideration that counts in favor of φ-ing.

Fortunately, this is one of the least contentious claims in metaethics. There is a lot of disagreement about the nature of the counts-in-favor-of relation; about its relata, its metaphysical underpinnings, and even how many places this relation has. But, despite this substantial disagreement, everyone who thinks that there are objective normative reasons agrees that they are considerations that count in favor of performing an action or adopting an attitude. Indeed, this claim is close to the status of a conceptual truth; someone who rejects it is simply not talking about what the rest of us are talking about when we talk about reasons.

Besides this basic claim, I need make no further assumptions about the counting-in-favor-of relation in this paper. So, to avoid making unnecessary enemies, I will make no further assumptions about it.

2. A simple argument

Why should we think that the moral rightness of an action is a reason to perform it?

Here is a simple argument:

1. A reason to perform an action is a consideration that counts in favor of performing it.
2. The moral rightness of an action is a consideration that counts in favor of performing it.
3. Therefore, the moral rightness of an action is a reason to perform it.

I do not think that this argument's simplicity is deceptive. I think that the argument is sound.

The argument's first premise is just a substitution instance of the universally-accepted claim, REASONS. And the argument is clearly valid. So the success of the argument turns on the truth of its second premise. In this paper I will therefore defend the argument's conclusion by defending this premise: the moral rightness of an action is a consideration that counts in favor of performing it.

I think this claim is so plausible that those who deny it bear the burden of proof. Consider the alternatives. There are two: that an action's moral rightness has no bearing at all on whether to perform it, and that an action's moral rightness counts against performing it. But both of these alternatives are *prima facie* absurd. Only in an amoralist's wildest dreams could it be that an action's moral *rightness* counts *against* performing it. And it is equally silly to think that an action's moral rightness has no bearing at all on whether to perform it. Perhaps morality is not overriding, in which case an action's moral rightness may not settle the question whether to perform it. But it is supposed to be part of the very concept of moral rightness that morally right actions are "to-be-performed" (on this see especially Mackie 1977, and the ensuing literature). So an action's rightness clearly has some bearing on whether to perform it, and its valence is clearly positive. Failing to understand this amounts to failing to understand what moral rightness is.

Here is another way of drawing out the *prima facie* plausibility of the simple argument's second premise. Imagine that you face a big, red, unmarked button. You have no idea what the effects of pressing it will be. At this point, you have no reason to press it, and no reason not to press it (except perhaps for reasons of curiosity). Then your favorite

omniscient, omnibenevolent, and trustworthy interlocutor appears and says "Look, you've really got to press this button as soon as possible. In fact, it's morally required of you to press it…" – at which point they unfortunately disappear again. I think it is very plausible that you now have some reason to press the button. So there has been a change to your normative situation: you went from having no reason to press the button to having some reason to press it. This change to your normative situation was occasioned by your learning that pressing the button is morally right. One simple and attractive explanation for this change is that the fact that you learned – the fact that pressing the button is morally right – is a reason to press the button.

Here is a final way of drawing out the *prima facie* plausibility of the simple argument's second premise. It is noteworthy that, when we learn of an agent's motivating reason for performing an action, we can then evaluate the quality of that motivating reason. We can ask whether it was a *good* reason for performing the action, or whether the action was *rational* or *reasonable*, or whether it *makes sense*. When we do this, we are looking for a match between the agent's motivating and normative reasons: we are asking ourselves whether her motivating reason for performing the action was a genuine normative reason to perform it. For example, encountering someone who goes around throwing pencils at people if she dislikes their smell, or who immediately rushes home to make a cup of tea whenever she sees a rabbit, doing so "because there was a rabbit", we say that her behavior is *irrational* or *unreasonable* or that her actions *make no sense*. We say this because her motivating reasons are not genuine normative reasons to perform these actions.

Now consider this adaptation of a fictional case:

> STAR WARS: Stormtrooper FN-2187 was bred and trained to fight for the evil First Order, the current incarnation of the dark side of the force. But, when

> sent on his first intergalactic mission, he is shocked and appalled by the blood spilt and the 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. FN-2187 chooses to rescue Poe and escape with him. On hearing 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*".

In STAR WARS, FN-2187's motivating reason (let's stipulate[14]) is the fact that his course of action is morally right. But he is not at all like the rabbit-tea-maker and the smell-pencil-thrower. Their behavior seems irrational, unreasonable, or even nonsensical. His makes perfect sense, and even seems pretty good. He seems to be doing a much better job at responding to genuine normative reasons than they are. One simple and attractive way to accommodate this intuition is to accept that FN-2187's motivating reason – the moral rightness of his course of action – is a genuine normative reason to take this course of action.

This is far from a knock-down argument for the claim that the moral rightness of an action is a consideration that counts in favor of performing it. There are ways for my opponents to wriggle out of each of the above attempts to shift the burden onto them: there are ways of construing "bearing" such that an act's rightness can be said to have some bearing on whether to perform the action, without counting in favor of performing it; there are alternative explanations of the change to your normative situation vis-à-vis the red button when you learn that button-pressing is morally required; and there are rival explanations

[14] I stipulate this to avoid addressing the fact that, as this dialogue continues in the actual movie, it is suggested that FN-2187 is mostly interested in using Poe's flying skills as a means of his own escape. I encourage the reader to consider an adaptation of the fictional case in which this does not hold, and FN-2187 is moved by conscience alone.

of FN-2187's apparent well-motivatedness. But I hope to have at least done enough to emphasize the *prima facie* plausibility of the claim that the moral rightness of an action is a reason to perform it.

3. We Can Have Our Buck and Pass It Too

a. The redundancy intuition

The main argument that has been offered against my thesis comes from those who accept a certain popular move in metanormative theory, applied to moral rightness. The popular move is called "buck-passing". In general, to "pass the buck" with respect to a moral property *M* is to make two claims about this property: first, that something's being *M* is not itself a reason for action, and second, that *M* is instead a status that something has in virtue of our (non-*M*) reasons for action.

For example, consider a curry. Suppose the curry has some properties: it is spicy, warm, and nourishing. And suppose that these properties make it the case that the curry has a further property: it is *good*. Perhaps it is good *qua* curry, or perhaps it is good absolutely. This distinction does not matter for present purposes. What matters is that a buck-passer about goodness will deny that the goodness of the curry is itself a reason to eat it. She will say that the lower-order, good-making features of the curry are reasons to eat it, and the resultant goodness of the curry is a status that the curry has *in virtue of* these features that are reasons to eat it, but is not *itself* a reason to eat the curry. This is a buck-passing view about the goodness of curries. (Buck-passing about goodness has been the subject of much discussion; for defenses, see e.g. Scanlon 1998, Parfit 2001, Olson 2004, Suikkanen 2004, Stratton-Lake and Hooker 2006, Skorupski 2007, and for criticisms see e.g.

Rabinowicz and Rønnow-Rasmussen 2004, Crisp 2005, Väyrynen 2006, Liao 2009, Gregory 2014). Views like this are called "buck-passing" because they "pass the normative buck" from goodness to the features of objects that their goodness consists in.

Buck-passing about rightness is a lot like buck-passing about goodness (for defenses see especially Dancy 2000, Stratton-Lake 2003; cf. Darwall 2010, Bedke 2011). Here's how it works. Consider an action. Suppose the action has some properties: it is fair, honest, and benevolent. And suppose that these properties make it the case that the action has a further property: it is *morally right*. A buck-passer about rightness will deny that this further property of the action is a reason to perform it. She will say that the lower-order, right-making features of the action are reasons to perform it, and that the resultant rightness of the action is a status that the action has *in virtue of* these features that are reasons to perform it, but is not *itself* a reason to perform it. This is a buck-passing view about the rightness of actions.

The main argument that buck-passers have offered for their view is based on the supposed *redundancy* of appeals to moral properties as reasons, once we acknowledge that the features that make it the case that these moral properties are instantiated are already reasons.

This began with buck-passing about goodness. Here is T. M. Scanlon (1998, p.97):

> [T]he natural properties that make a thing good or valuable… provide a complete explanation of the reasons we have for reacting in these ways to things that are good or valuable. It is not clear what further work could be done by special reason-providing properties of goodness and value.

Consider again the curry. Scanlon's thought can be expressed by saying that the curry's spiciness, warmth, and nourishing-ness surely provide reason enough to eat it, without the resultant goodness of the curry also being a reason to eat it. To Scanlon, it seems that there is simply no point in saying that the curry's goodness is also a reason; it seems to make no difference to an agent's normative situation, once the more fundamental properties that explain its instantiation have already been taken into account. This is the idea that Scanlon expresses by saying that "it is not clear what further work could be done" by goodness.

Scanlon's critics have pointed out that parallel remarks apply to Scanlon's own view that the rightness of an action is a reason to perform it. Here, for instance, is Philip Stratton-Lake (2002, p.15):

> I can see no reason why [we] cannot understand rightness as well as goodness in terms of reasons. [We] could (and in my view should) embrace not only a buck-passing account of goodness, but also a buck-passing account of rightness. According to such an account, the fact that φ-ing is right is the same as the fact that φ-ing has properties that give us conclusive reason to do it. Similarly, the fact that φ-ing is wrong is the same as the fact that it has properties that give us conclusive reason not to do it.

Stratton-Lake applies this line of thought as a criticism of Scanlon's contractualism in his (2003). But the problem is not unique to Scanlon. Stratton-Lake's general thought can be expressed as follows: the fact that an action is morally right is not a brute fact. It is rather a fact that obtains in virtue of further features of the action. Since they make this action morally right, these features must give us conclusive reason to perform it. But then it seems as though there's no point in saying that the action's rightness provides a further reason to perform it, on top of the reasons provided by the right-making features – by stipulation, those reasons were already conclusive! So the action's rightness seems to

make no difference to an agent's normative situation, once the more fundamental properties that explain its instantiation have already been taken into account. It is on these grounds that Stratton-Lake suggests that we identify the fact that an act is right with the fact that it has properties that give us conclusive reason to do it, rather than seeing it as a further fact that may be a reason in its own right.

I will use the phrase "the redundancy intuition" to refer to the intuition that it is redundant to say that a moral property is a reason, once we are already taking the lower-order features that make it the case that the moral property is instantiated to be reasons. And I will use the phrase "the redundancy argument" to refer to the argument that we should not take a moral property to be a reason, on the grounds that doing so would elicit the redundancy intuition.

I have also heard it said that it is not only redundant but positively *inappropriate* to regard moral properties like goodness and rightness as reasons, in addition to the good- or right-making features. It has been suggested to me in conversation that this seems like an illegitimate form of double-counting. This thought is surely closely related to the redundancy intuition; in what follows I will construe it as a species of the redundancy intuition.

The redundancy argument is the main motivation for buck-passing about rightness. But this argument rests on a mistake. The redundancy intuition is picking up on something important: an action's rightness does not always add *extra normative weight* in favor of performing it, no matter what else has already been taken into account. But this does not show that an action's rightness is not a reason to perform it. Rather, it shows that reasons for action do not always add extra normative weight in favor of the actions for which they are reasons, no matter what else has been taken into account. So I will now argue.

b. The problem: rampant redundancy

The redundancy argument rests on a mistake. We can see this by noting that the argument overgeneralizes. It is possible to elicit the redundancy intuition about all sorts of properties, some of which very plausibly are reasons for action. The redundancy argument then applies to these features. Ultimately, the argument suggests that no fact that is metaphysically constituted, wholly or partly, by another fact that is a reason to φ can itself be a reason to φ. But the idea that some facts of this kind are reasons is considerably more plausible than the redundancy argument itself.

We can begin to illustrate this point by looking at the very features of actions that buck-passers champion: the right-making features. These are supposed to be reasons for action, according to buck-passers. But it is possible to elicit redundancy intuitions about these features. That is because the right-making features are not fundamental. Facts about the instantiation of these features are not brute facts, insusceptible of further explanation. Rather, when an action possesses a right-making feature, it does so in virtue of further, lower-order features of the action. Many of these further features seem like great reasons to perform the action. But, if the lower-order features are already reasons, then it is just *redundant* to suppose that right-making features are also reasons. So the redundancy argument applies to the right-making features as well.

For example, suppose that fairness is a right-making feature. The fact that an action is fair is not plausibly a brute fact, insusceptible of further explanation. On the contrary, it is constituted by further facts; perhaps the fact that the action distributes social benefits and burdens on reasonable, non-arbitrary grounds. These are in turn constituted by further facts; perhaps the fact that the action is meritocratic, or that it makes reparations for past

injustice, or that it gives to those with the highest need. These lower-order facts all seem like great reasons to perform the action. But if we take any of them to be reasons, then it seems *redundant* to do the same for the higher-order fact that the action is fair. The redundancy argument then holds that, since it is redundant to take the action's fairness to be a reason to perform it, we should not do so. But the action's fairness is a right-making feature. So the redundancy argument excludes right-making features as potential reasons, just as much as it excludes rightness itself.

This example makes trouble for buck-passers because they hope to establish *both* that an action's rightness is not a reason to perform it *and* that our reasons for action are the right-making features. It will be difficult for them to accomplish both of these aims, because the redundancy argument – buck-passers' strategy for denying that rightness is a reason – undermines their positive claim that the right-making features are reasons.

To make matters worse, the redundancy argument is not even limited to cases involving moral properties (like rightness and fairness). Here is a non-moral example:

> 5-A-DAY: In the UK in the early 2000s, there was a public health campaign to get people to eat at least five portions of fruit or vegetables each day. As a result, supermarkets now put stickers on their prepared food that say "1 of your 5-a-day!", "2 of your 5-a-day!", and so on. Supermarkets produce stickers reporting the number of portions of fruit or vegetables in their food; if a salad contains, e.g., three portions of vegetables, then it is labeled with a single "3 of your 5-a-day!" sticker rather than three "1 of your 5-a-day!" stickers. But supermarkets do not produce stickers naming the particular fruits or vegetables in their food; a snack pack containing one portion of

apple will have a "1 of your 5-a-day!" sticker, rather than a "1 portion of apple!" sticker. The latter are not manufactured.

If the redundancy argument is correct, then British supermarkets' behavior in 5-A-DAY is quite mysterious. They label a salad containing three portions of vegetables with a single "3 of your 5-a-day" sticker, as if this were a reason to eat the salad. But the fact that a salad contains three of your 5-a-day consists in the fact that it contains one of your 5-a-day, and then one more one, and then one *more* one, since this is what it is to contain three of something. And this lower-order fact is surely a perfectly good reason to eat the salad. So why produce stickers mentioning any number of portions greater than 1? Isn't this *redundant*, if we already have multiple "1 of your 5-a-day" stickers? Similarly, a salad's containing three of your 5-a-day may consist (for example) in its containing one portion each of lettuce, tomato and cucumber. And the salad's containing these three vegetables surely counts in favor of eating it. (Indeed, it is hard to see how the salad's containing three of your 5-a-day could count in favor of eating it if containing lettuce, tomato and cucumber *doesn't* count in favor of eating it, and these are the vegetables that the salad contains.) But, then, isn't it *redundant* to say that the fact that the salad contains three of your 5-a-day is a reason for eating it? If lower-order facts about particular vegetables are already reasons, then why should we count higher-order facts about portion numbers as reasons to do anything at all?

Examples like this show that the redundancy intuition generalizes. The intuition arises whenever one fact that counts in favor of performing some act is made the case by a further fact or facts, at least some of which also count in favor of performing the same act. So the redundancy argument, if correct, applies in all these cases too. But it is just not plausible that *none* of the higher-order facts in *any* of these metaphysical hierarchies are reasons. So the redundancy argument overgeneralizes. Redundancy is rampant.

c. *The solution: The Buck Doesn't Stop Anywhere*

Something must be wrong with the redundancy argument. What is it?

Here is what I think is wrong. The redundancy argument identifies a metaphysical hierarchy wherein one fact that counts in favor of φ-ing is metaphysically constituted by further facts, at least some of which also count in favor of φ-ing (and may themselves be constituted by further facts, some of which also count in favor of φ-ing). The argument then assumes that, for each such hierarchy, there must be a privileged level at which the reason "really" lies. The metaphor of "buck-passing" unhelpfully encourages this idea; this metaphor conjures up an image of the reason – the "buck" – being passed down from the less to the more fundamental levels in a metaphysical hierarchy, until at some point the music stops and one lucky fact is left holding the buck. This picture suggests that if we just examine these metaphysical hierarchies carefully enough, then we will eventually be able to identify a special fact in each one that is the "real" reason. On this way of thinking, identifying genuine normative reasons is like spotting Waldo in a crowd.

I think that this approach is silly. We need not locate the normative buck at any particular level in a metaphysical hierarchy. Rather, we can and should say that the facts in these hierarchies can all be reasons.

This point is especially easy to see on some metaphysical suppositions about the relationship between right-making features and rightness. Suppose that there is just one right-making feature (the maximization of value, perhaps) and that rightness is type-identical to this feature. Or suppose that there are multiple right-making features, but that each instance of moral rightness is token-identical to an instance of one of these

features. On either of these metaphysical pictures, the fact that an action is morally right wholly consists in some fact, *P*, about the action's instantiation of a right-making feature – the fact that *P* and the fact that the action is morally right are literally the same fact. In this case there is no question of where to locate the buck, and no good grounds for a redundancy argument. There is a sense in which the redundancy intuition is correct: counting both the action's moral rightness and the fact that *P* as reasons to perform the act would literally be double-counting. But this hardly shows that the action's moral rightness is not a reason to perform it. On the contrary, it shows that the action's moral rightness *is* a reason to perform it: the reason that *P*. In short, if rightness is either type- or token-identical to the right-making feature/s, then we cannot maintain that the right-making features are reasons but rightness is not, as this violates the indiscernibility of identicals.[15] On either of these metaphysical pictures, buck-passing about rightness is simply incoherent.

Buck-passing becomes a viable option if the relationship between rightness and the right-making features is not any kind of identity, but some more complicated relationship like metaphysical grounding. On one of these more complicated pictures, it is not logically true that rightness is a reason iff the right-making features are. So there is a live question as to whether it is rightness *or* the right-making features *or* both that are reasons.

But even on this metaphysical picture we need not identify a particular point in the metaphysical hierarchy at which to locate the normative buck. Distinguish two possible views:

[15] This assumes that the predicate "… is a reason to ϕ" is not hyperintensional (thanks to Umer Shaikh and Pekka Väyrynen for pointing this out to me). When we are talking about objective normative reasons, I find this assumption plausible. For example, if pain is identical to a certain brain state, then the fact that a patient is in pain is an objective normative reason to administer medication iff the fact that the patient is in the brain state is an objective normative reason to administer the medication. Since there is just one fact here, it either counts in favor of performing an action or it doesn't – regardless of whether anyone is *motivated* by thoughts about it under a particular description.

> "SPECIAL FACT" VIEW: In a metaphysical hierarchy in which some facts that seem to count in favor of performing an act are metaphysically constituted by others that seem to count in favor of performing the same act, there is always one fact that is *where the buck stops* – a special fact that bears all the normative weight.
>
> "SHARE THE WEIGHT" VIEW: In a metaphysical hierarchy in which some facts that seem to count in favor of performing an act are metaphysically constituted by others that seem to count in favor of performing the same act, it can be that all of the facts that seem to count in favor of performing the act genuinely do count in favor of performing it. *The buck doesn't stop anywhere*. The normative weight is shared by all of the facts in the hierarchy rather than resting on some particular fact.

Buck-passers favor the "special fact" view, and they think that facts about right-making features are among the special facts. But, given that the redundancy argument dramatically overgeneralizes, the "share the weight" view is the more attractive option. Faced with metaphysical hierarchies of facts that each seem to count in favor of performing some action, we should abandon the project of examining each hierarchy to see where in it we can discern a "buck" nestling on a special fact. Instead, we should embrace the possibility that most or even all of the facts that seem to count in favor of performing the action really do count in favor of performing it, and thus are genuine objective normative reasons to perform it.

If we adopt the "share the weight" view, then we can say that *any* fact that counts in favor of performing an action is a reason to perform it – including, for example, the fact that

the action is morally right, the fact that it is fair, the fact that it distributes benefits and burdens on reasonable, non-arbitrary grounds, and the fact that it is meritocratic, makes reparations for past injustice, and/or gives to those with the highest need. The "share the weight" view thus preserves one of the few claims in metanormative theory that enjoys widespread consensus: the claim that a reason to φ is a consideration that counts in favor of φ-ing. I take this to be a significant benefit of the view.

Let me forestall a possible misunderstanding, which would be a substantial misunderstanding (and which is, I think, the misunderstanding underlying the redundancy intuition). When I say that multiple facts in a single metaphysical hierarchy might all be genuine normative reasons with shared weight, I do not mean that these facts are all always equally appropriate to cite as reasons when deliberating about what to do, or when evaluating our own or others' decisions. This is not true. On the contrary, when multiple facts in a single hierarchy are reasons with shared normative weight, features of our conversational context often determine whether it is more appropriate to cite the lower-order or the higher-order facts as reasons. There are often substantial differences between the complete picture of *what counts in favor* of performing a certain action and *what we should count* as favoring the action in our conversational context. Sometimes it is more appropriate to be succinct, focusing on the higher-order facts. And at other times it is more appropriate to be detailed, focusing on the lower-order facts.

This is what makes sense of supermarkets' behavior in in the 5-A-DAY example. Supermarkets are not using mere shorthand or metaphor, nor are they making a mistake, in using a single "3 of your 5-a-day!" sticker rather than three "1 of your 5-a-day!" stickers, purportedly alerting customers to a reason to eat a salad. The fact that the salad contains three of your 5-a-day *is* a reason to eat it. And a single "3 of your 5-a-day!" sticker conveys the same information as three "1 of your 5-a-day!" stickers, but does so more

succinctly. Similarly, supermarkets' decision to report numbers of portions, but not particular fruits or vegetables, also makes sense in context. British consumers are encouraged to count portions in order to achieve the 5-a-day goal, ignoring the nature of the particular fruits or vegetables comprising those portions. This makes it more appropriate to notify them of the number of portions in a salad than the particular fruits or vegetables.

The potentially relevant features of conversational context here are as many and as varied as features of conversational context usually are. Some features favor detail over summary, making it more appropriate to cite lower-order facts as reasons. For example, if someone has a tomato craving, it might be appropriate to think of her reason to eat the salad as the fact that it contains *tomato*, since this fact is particularly relevant to her interests. For another example, if someone is choosing between two 5-a-day snack packs that each contain five portions of fruit and vegetables, then it might be appropriate to cite the particular fruits and vegetables in each snack pack (and/or their nutritional properties) as her reasons to eat them, as this is what distinguishes between the two options. When we are interested in comparing the reasons favoring one act with the reasons favoring another, it is helpful to cite facts in the metaphysical hierarchies favoring each act at a level of generality that distinguishes between them.

Other features of conversational context favor summary over detail, making it more appropriate to cite higher-order facts as reasons. For example, suppose that someone knows that she prudentially ought to eat healthily, but is tempted to eat a delicious yet ludicrously unhealthy chip butty for lunch. In this context it is appropriate to get across the difficulty of her decision by saying, "Well, a salad contains three of her 5-a day, but a chip butty sure is tasty!". This succinctly conveys the salient difference between her lunch options. Describing the vegetables in the salad would add unnecessary detail.

For another example, suppose that someone reading a menu sees that a salad is described as containing "a medley of seasonal vegetables dressed with a tangy house vinaigrette". In this context it is appropriate to cite the fact that the salad contains *vegetables*, rather than any facts about the vegetables it contains, as the agent's reason to order it. That is because only the former fact could be the agent's *motivating* reason to order the salad, since this is the only fact to which she has epistemic access. Often, when we identify an agent's normative reasons, we are looking for a match between normative and motivating reasons – the sort of match that is absent in the cases of the smell-pencil-thrower and rabbit-tea-drinker in §2. But identifying reasons too low-down in a hierarchy can create a false sense of mismatch. If the agent orders the salad because it contains vegetables, but we say that her "real" normative reason to order it is that it contains lettuce, cucumber and tomato, then we make it look as though her motivating reason is not a genuine normative reason, when in reality it is a perfectly good normative reason that is simply less fundamental than the fact we chose to mention.

These observations all suggest that, even when the relationship between facts in a hierarchy is some sort of metaphysical constitution that falls short of identity, we should not expect there to be a privileged level at which the reason "really" lies. Rather, context determines which reasons it is appropriate to consider in each case. At this point, the other facts in the hierarchy – whichever they are – all begin to seem redundant. But this kind of redundancy cuts both ways; either lower-order or higher-order facts can be made to seem redundant by the salience of other reasons in the same hierarchy. If the important thing, in a context, is that a salad has three portions in it, then it is redundant to ruminate on the specific vegetables once the fact that there are three of them has already been noted. Likewise, if the important thing in a context is that a salad contains tomato, then

it is redundant to mention the fact that it contains a vegetable (i.e. tomato) once the presence of tomato has already been noted.

Most of the points in this discussion of salads have direct analogues when it comes to moral rightness. The features of conversational context that favor summary over detail can all favor citing an action's rightness as the agent's reason to perform it, rather than spelling out the right-making features. Consider the STAR WARS case again. We convey FN-2187's heroism at a level of abstraction suitable for most conversational contexts by saying that he faced a choice between doing what's right and doing what's easy, and he chose to do what's right. As in the case of the salad and the chip butty, this draws a contrast between FN-2187's moral and prudential reasons for action in a way that avoids unnecessary detail.

For an analogy with the case of the menu, consider someone who remembers or is told that a certain course of action is morally right, but does not remember or is not told what its right-making features are. Perhaps she has reasoned her way to the conclusion that maintaining a vegan diet is morally required on multiple long dark nights of the soul, and is thus confident that maintaining a vegan diet is morally required, though she cannot recall the subtleties of her reasoning. Or perhaps she consults an expert on inclusive pedagogy to determine the right response to the fact that fliers containing racist messages have been posted all over her campus and she now must teach her undergraduate class, and the expert tells her what to do but does not have time to explain the right-making features. In either of these cases the agent may choose to take a certain course of action *because it's the right thing to do*, although she is not in a position even to say what its right-making features are, let alone to be motivated by them – she lacks epistemic access to the relevant facts. If we say that these agents' "real" normative reasons

are the right-making features, then we create precisely the same false sense of mismatch between motivating and normative reasons as in the menu case.

Unsurprisingly, there are some disanalogies between salads and moral rightness. One is that it is possible for there to be many salads that all contain three portions of vegetables, whereas it may be (if there are no genuine moral dilemmas) that at most one act can be morally required of an agent at any one time. Another disanalogy is that, though an agent's tomato craving can make it conversationally relevant that a salad contains *tomato*, rather than any other fruit or vegetable, the analogous view seems badly mistaken when it comes to morality. If someone cares which lower-order facts make an action right – for example, because she is indifferent to considerations of well-being but cares deeply about justice – it seems mistaken to think that this could be relevant to her objective normative reasons for action. Whether an action's rightness, well-being-promotion, or justice is an objective normative reason for an agent to perform it does not depend on how much she personally likes rightness, well-being, or justice. In this respect, justice is unlike a tomato.

But these two disanalogies just eliminate two ways in which, when it comes to morality, it might be more appropriate to cite lower-order facts as reasons than it is to cite higher-order facts. Two ways for context to make it more appropriate to consider the more fundamental facts when examining an agent's reasons for action cannot arise when it comes to moral rightness. So we should expect there to be plenty of contexts in which the moral rightness of an action is the salient reason to perform it, and in which further consideration of its right-making features as reasons is thereby rendered redundant.

d. A problem for the buck-passer

The above reflections highlight a problem for the buck-passer. Consider again the agents who remember or are told that a certain course of action is morally right, and who choose to undertake the course of action on this basis, but who do not remember or are not told what its right-making features are. One agent has reasoned her way to the conclusion that maintaining a vegan diet is morally required on multiple long dark nights of the soul, and is thus confident that maintaining a vegan diet is morally required, though she is unable to recall the subtleties of her reasoning. Another agent consults an expert on inclusive pedagogy to determine the right response to the fact that fliers containing racist messages have been posted all over her campus and she now must teach her undergraduate class, and the expert tells her what to do but does not have time to explain the right-making features. Here is a plausible claim about these agents: in taking the courses of action that they know to be morally right, and taking them on this basis, they are doing a better job of responding to reasons than the agent who makes cups of tea "because there was a rabbit".

I take this to be a very plausible claim. The case of the rabbit-tea-maker elicits the sense of mismatch between an agent's motivating and normative reasons that is characteristic of irrational actions that make no sense, while the cases of partial forgetting and testimony do not. But it is difficult for the buck-passer to account for this. The agents in these cases ostensibly choose to take a course of action *because it's the right thing to do* – as does Stormtrooper F1-287 in STAR WARS. So it seems fair to assume, in each case, that the action's rightness is the agent's motivating reason. But the buck-passer denies that this can be a normative reason. So the buck-passer is committed to saying that these cases display the mismatch between normative and motivating reasons that is characteristic of irrationality. This seems unduly harsh.

The buck-passer has some options that may help her to avoid this verdict in some cases, but none (so far as I can see) that work well in all cases. Here I will survey three options and some limitations of each of them.

The buck-passer might claim that Stormtrooper FN-2187 – contrary to his claims – is not really motivated by the rightness of his act. She can say that, when he says he is helping Poe to escape "because it's the right thing to do", this is just an elliptical way of referring to the right-making features, which are what *really* motivate him. FN-2187 presumably does have some grasp on the right-making features, so they might be what motivates him. This would secure the desired match between FN-2187's motivating and normative reasons; they are both, on this account, the right-making features of his act.

But this strategy does not work for cases of partial forgetting and testimony. In these cases, one agent has forgotten the right-making features, and another is completely unaware of them. So if the buck-passer wants to say that facts about the right-making features are these agents' motivating reasons, then she has to say that one agent's motivating reasons are facts that she has forgotten, and another agent's motivating reasons are facts of which she is wholly unaware. This is an extremely bizarre view of motivation. It is unproblematic to say that the right-making features are reasons for these agents *to* act. But it is quite odd to say that they are the reasons *for which* the agents are currently acting – the considerations that are moving them to act – given that the agents themselves could not possibly have any idea that this is the case, not because the relevant facts are buried deep within their subconscious but because they have *no epistemic access whatsoever* to the relevant facts. It is very hard to see how someone could be motivated to act by a fact on which she has no epistemic grasp whatsoever. So, this strategy requires us to construe these agents as objectionably alienated from the reasons on which they are currently acting.

Here is a second option. The buck-passer can create a match between motivating and normative reasons in cases of partial forgetting and testimony by holding that the relevant facts are not facts about the actions' rightness, nor facts about their right-making features, but rather some facts about the agents' evidence, or about their doxastic states. For example, the buck-passer could suggest that in cases of partial forgetting the agent's normative and motivating reason is that she *seems to remember* that sticking to a vegan diet is morally right, and in cases of testimony the agent's normative and motivating reason is that she *was told by a moral expert* that the intervention with her students is morally right. Or the buck-passer could suggest that the reason in each case is that the agent *believes* that the act is right. We do sometimes cite an agent's evidence or doxastic states as her reason to act, especially when her evidence is misleading. For instance, if my evidence suggests that it will rain tomorrow, but it won't rain tomorrow, then we say that the fact that I *believe* that it will rain, or the fact that *my evidence suggests* that it will rain, is a reason for me to bring an umbrella. (We can't say that the "fact" that it will rain is a reason for me to bring an umbrella, since this isn't a fact.) So this is another way for the buck-passer to go.

But this, too, is an odd take on the cases. It is noteworthy that we usually cite facts about an agent's beliefs or evidence as normative reasons *only* when the evidence is misleading. When someone comes to know a fact as a result of her evidence, we usually say that this fact is her reason, rather than a fact about her beliefs or evidence. For example, if I can see tomato in a salad, we usually say that my reason to eat the salad is that *it contains tomato*, not merely that *it seems to contain tomato* or that *I believe it contains tomato*. And in the cases of partial forgetting and testimony the agents' evidence is not misleading. So it is odd for the buck-passer to say that their normative reason is merely some fact about their evidence or doxastic states, rather than the fact that they come to know as a result of this

evidence: the fact that their acts are morally right. Moreover, this is even odder in FN-2187's case. We may suppose that his knowledge that helping Poe to escape is morally right is as robust as can be. In light of this, it seems unduly skeptical to insist that his normative reason to help Poe to escape is just that *he believes* or *his evidence suggests* that it is morally right.

Here is a final option. The buck-passer could argue that agents' normative reasons to act in cases of partial forgetting or testimony are not the rightness of their actions, but some further fact that is entailed by the actions' rightness. For instance, the fact that an action is right entails that it has at least one right-making feature, and that the balance of moral reasons favors performing it. So the buck-passer could say that one of these facts is a normative reason to perform the action. This lets the buck-passer preserve the intuition that someone who learns or remembers only that an action is morally right already has epistemic access to a fact that is a normative reason to perform the act, while denying that the action's moral rightness is itself a reason to perform it.

This seems like a desperate move. While it does allow the buck-passer to continue to deny that rightness is a reason, it does so only at the expense of allowing that another metaphysically higher-order property – the property of having at least one right-making feature, or the property of being favored by the balance of moral reasons – is a normative reason. But the whole point of buck-passing was supposed to be to *avoid* citing metaphysically higher-order properties as reasons, since this supposedly leads to redundancy. Since the buck-passer holds that right-making features are reasons, allowing another higher-order property to be a reason will raise precisely the same worries about double-counting and redundancy that were the basis of her initial denial that rightness is a reason. If she accepts the "special fact" view and the redundancy argument, then she cannot say both that (e.g.) the fact that an act is fair and the fact that it has at least one

right-making feature are reasons to perform it. These facts occur in a single metaphysical hierarchy, so they cannot both be the special fact on which the buck nestles. And mentioning both together invokes the specter of redundancy. So this option lands the buck-passer right back where she started.

Buck-passers can hold out hope and try to find a more plausible thing to say about these cases. But there's no point in doing this. The main argument against saying that an action's moral rightness is a reason to perform it rests on a mistake. So, we should go ahead and say, *pace* buck-passers, that an act's moral rightness is indeed a reason to perform it. That easily takes care of these cases.

e. How to understand the redundancy intuition

I have suggested that features of our conversational context often determine which facts in a metaphysical hierarchy it is most appropriate to mention when discussing an agent's reasons for action. This helps us to understand the redundancy intuition. The redundancy intuition does not tell us what is a "real" reason and what is not – it does not tell us how to identify or individuate reasons. But it might tell us something about how to aggregate an agent's reasons for action, given the metaphysical relationships between them, to determine the *total amount of normative weight* favoring the performance of a certain act.

In a conversational context, when we list an agent's reasons for action, we are typically trying to determine the total amount of normative weight favoring the agent's performance of each action. This may be because we want to compare the total amount of normative weight favoring one action to that favoring an alternative: we want to know what the agent has *most reason* to do. Or it may be because we want to see whether the total amount of normative weight favoring the performance of a certain action passes a

certain threshold: we want to know whether there is *sufficient reason* to perform it. This means that citing multiple facts in a single metaphysical hierarchy can be misleading. When facts share their normative weight, citing many of them can make it seem as though the total amount of normative weight favoring one action is greater than it really is. By analogy, to say "this salad contains three of your 5-a-day, *and* it contains lettuce, cucumber, and tomato!" misleadingly suggests that the salad contains six vegetables in total. I see redundancy intuitions as warnings that two or more reasons occur in a metaphysical hierarchy with shared weight, and thus that mentioning all of them may misrepresent the total amount of normative weight favoring an action.

This clarifies the sense in which redundancy cuts both ways. In describing an agent's reasons to perform an action, we should say enough to convey the total amount of normative weight favoring performing the action, and no more, and no less. It is often the case that mentioning (all) the facts at *any* level in a metaphysical hierarchy would be sufficient to convey the total amount of normative weight that these facts share, and we simply have to pick a level. But, once we have mentioned the reasons at one level, the others become redundant to mention, as their normative weight is already accounted for.

The task of saying enough to convey the total amount of normative weight favoring the performance of an action, no more, and no less, can be complicated by the complicated structure of normative reality. There are cases in which it is *not* redundant to mention two or more of the facts in a single metaphysical hierarchy, because not all their normative weight is shared. For example, recall the agent who has a tomato craving. In describing the reasons for her to eat a salad, we can count both the fact that it contains three of her 5-a-day and the fact that it contains tomato as reasons to eat it, even though tomato is one of the three portions. This is a sort of double-counting. But it is not objectionable, because it accurately reflects normative reality. The fact that the salad

contains tomato really does count "doubly" in favor of eating it: as a way of moving toward the 5-a-day goal, and as a way of satisfying the agent's tomato craving. (We might convey this by saying "This salad contains three of her 5-a-day – *and* one of them is *tomato*!") The lesson to draw here is that a single fact may figure in multiple metaphysical hierarchies that each bear normative weight favoring the performance of a certain action. If the goal is to convey the total amount of normative weight favoring performing the action, and no more, and no less, then we *should* count this fact multiple times.[16]

The phenomena here are pragmatic rather than metaphysical. Concerns about how to felicitously convey the amount of total normative weight favoring an act do not tell us that only some of the facts that we might mention are "real" reasons. We can tell that the phenomena are merely pragmatic by observing that the most appropriate reasons to mention vary across conversational contexts – as they do in response to considerations favoring succinctness over detail, or *vice versa*. (They also vary in response to audiences' prior knowledge; for instance, saying that a salad contains three of your 5-a-day is unilluminating to someone who has never heard of the 5-a-day campaign.) Further evidence comes from the observation that it is possible to cite two or more facts with shared normative weight *without* eliciting the redundancy intuition, if we use the right verbal cues. For instance, it is fine to say "this salad contains three of your 5-a-day – *namely*, lettuce, cucumber and tomato", or "this act is morally right *insofar as* it is fair – *that is to say*, it is meritocratic and makes reparations for past injustice". By using verbal cues like "namely", "insofar as", "that is to say", "in virtue of", and so on, we indicate that we are describing facts in a single metaphysical hierarchy with shared normative weight. This is not redundant, and is sometimes positively helpful, as it gives the audience more information about the structure of normative reality.

[16] Thanks to Matt Bedke, Gunnar Björnsson, Justin Snedegar, and Daniel Wodak for helpful discussion of this point.

There is a kernel of metaphysical truth behind the redundancy intuition. The kernel of truth is that some reasons share their normative weight with facts that they constitute or by which they are constituted. This means that not all reasons always add extra normative weight favoring the acts for which they are reasons. Whether a reason adds weight depends on what other reasons we have already taken into account. If a reason shares its normative weight with others that we have already taken into account, then it has nothing to add. This is an important insight about how to aggregate reasons. But, though important, this insight is not groundbreaking. We already knew that not all reasons always add further normative weight favoring the performance of actions for which they are reasons, no matter what else we have taken into account. We knew this from every example supporting holism or particularism, from every example of undercutting defeat, and from every example of combinatorial effects between reasons (see e.g. Schroeder 2009, Horty 2012, Nair 2016). We also knew this from the idea of exclusionary reasons – reasons such that their obtaining is itself a reason not to take other reasons to provide any normative weight (see Raz 1990). The kernel of truth behind the redundancy intuition is that there is another, underexplored, class of combinatorial effects arising from metaphysical relationships between reasons. But combinatorial effects do not in general imply that some apparent reasons are not "really" reasons. And they should not be taken to do so here.

f. Comparison with other literatures

The view about normative reasons that I have defended here has parallels with views on other topics in metaphysics and in the philosophy of science. For instance, consider mereological composition. Nobody is surprised or confused to learn that a statue weighing 200lb and a lump of clay weighing 200lb, laid together on a scale, weigh only

200lb. Once we understand that the statue and the clay *share their physical weight* – whether or not they are identical, and indeed whatever the metaphysical relationship between them turns out to be – there is no great mystery here. Similarly, nobody is surprised or confused to learn that an area within physical space is fully occupied both by a whole and by its parts. Here we do not think that double-counting intuitions suggest that at most one of the whole and the sum of the parts is "really" in the space. Likewise, I think, intuitions about the double-counting of shared normative weight do not suggest that at most one of the facts that share the weight is a real reason.

Stephen Yablo (1992) offers an argument, similar to my argument in §3b, for the view that the causal sufficiency of an event x need not exclude the causal relevance of another event x^* to a third event y, if x and x^* are appropriately metaphysically related to one another. Yablo focuses on the determinate-determinable relation, which is one plausible candidate for being the relationship between moral rightness and the right-making features. Here is one of his examples (*ibid.*, pp.257):

> Imagine a pigeon, Sophie, conditioned to peck at red to the exclusion of other colors; a red triangle is presented, and Sophie pecks. Most people would say that the redness was causally relevant to her pecking, even that this was a paradigm case of causal relevance. But wait! I forgot to mention that the triangle in question was a specific shade of red: *scarlet*. Assuming that the scarlet was causally sufficient for the pecking, we can conclude by the exclusion principle that every *other* property was irrelevant.

On Yablo's view, intuitions about double-counting (or, in the case of causation, about overdetermination) are simply confused in cases involving metaphysical relationships between causes, like the relationship between a triangle's being red and its being scarlet. I agree. I think the same thing about normative reasons.

The distinction I have drawn between *what counts in favor* of performing an act and *what we should count* in a conversational context also has parallels with existing views on causation and explanation. On one view, causation is a "broad and nondiscriminating" relation between events and their entire causal histories (cf. e.g. Bennett 1988, Lewis 2000), and there are no special events that are the "real" causes. But features of conversational context determine which past events it is most appropriate to cite as causes of a certain effect. Swanson (2010) discusses one way of spelling out this view, on which causal talk is governed by a principle enjoining us to use "good representatives" of the causal paths to an effect. This principle means that, once one event on a causal path to the effect has been mentioned, it can become infelicitous to mention other events on the same causal path, since mentioning multiple events as causes of an effect typically indicates that they are on distinct causal paths. Fogal (2017) offers an extensive comparison between the data about causation presented by Swanson and some data about normative reasons that he presents. On Fogal's account, similar pragmatic principles enjoin us to use good representatives of a "normative cluster" favoring an act, where a "cluster" may include items of different metaphysical types (e.g. an event and the fact of its occurrence) or facts that count in favor of an action taken collectively but not severally (e.g. the facts that there is dancing at a party and that Billy enjoys dancing). I agree. I think that parallel phenomena occur with respect to sets of facts arranged in relationships of metaphysical constitution.[17]

Similarly, on one view about explanation, the full explanation of a fact or event would be a maximally detailed account of absolutely everything at all relevant to the fact's obtaining or the event's occurring. But individual statements can still be explanatory, insofar as they tell us something relevant to our interests about what this full explanation

[17] I am grateful to both Swanson and Fogal for helpful discussion of the parallels between their accounts and the position that I develop in this paper.

would be like. Railton (1981) discusses one way of spelling out this view: he envisions an "ideal explanatory text" offering the complete account of some phenomenon, and notes that, in practice, we never need to know the entire content of this text. Rather, we are interested in learning information that accurately reduces our uncertainty about the content of the ideal explanatory text, to a degree and in a manner that is appropriate in our context. Railton says the following as to why good explanations may be partial (*ibid*., p.239):

> In certain contexts, a more elaborate explanation may be out of place – the audience may be too well-versed, not well-versed enough, not interested enough, or short on time; a more elaborate explanation may not be available even if it were appropriate – the relevant laws and facts may not be known, or may be known only qualitatively; the person offering the explanation may simply not know enough; and so on.

I agree. And I think that this provides a useful parallel with the way in which context determines which of the facts in a metaphysical hierarchy it is most appropriate to mention when determining the amount of total normative weight favoring performing an action.[18] We might imagine an "ideal normative text" that gives the full account of every fact bearing some normative weight that favors the action, and of the metaphysical relationships between them. But, in practice, we are rarely interested in coming to know the entire content of this text. (Perhaps those who study first-order ethical theory aim to discover the full content of the text – but ordinary agents do not.) Rather, we are interested in learning information that will accurately reduce our uncertainty about the total amount of normative weight favoring an action, and about how this amount compares to the amounts of normative weight favoring the alternatives.

[18] Thanks to Jim Joyce for helpful discussion of this point.

I find these parallels encouraging. They suggest that I am along the right lines in the diagnosis that I have offered of the mistake underlying the redundancy argument, and in the alternative picture that I have begun to sketch. Analogous mistakes have already been corrected in other branches of philosophy. It is time to correct this one.

4. Upshot

Here is a summary of what happened in this paper. I sketched a picture of moral metaphysics on which every fact that counts in favor of performing an act is a genuine normative reason to perform the act, notwithstanding the observation that reasons sometimes arise in metaphysical hierarchies with shared normative weight. I suggested that this can include both an action's rightness and its right-making features (and the features that make it the case that it has the right-making features, and the features that make it the case that it has *those* features, and so on). So an action's moral rightness and its right-making features can all be genuine normative reasons to perform it. We can have our buck and pass it, too.

I suggested that features of our conversational context often determine which of the facts in a metaphysical hierarchy it is most appropriate to mention when discussing an agent's reasons for action. But, I argued, it would be a mistake to think that this shows that the facts it would be less appropriate to mention are not really reasons. Pragmatic principles place constraints on how many and which of the normative reasons in a metaphysical hierarchy it is most appropriate to mention, consistent with the general conversational goal of conveying the total amount of normative weight favoring an action, no more, and no less. But this is a point about how to describe and aggregate reasons within a conversational context, rather than a point about which sorts of facts can be reasons in

the first place. And there are plenty of contexts in which pragmatic mechanisms ensure that the moral rightness of an action is the salient reason to perform it.

In short: there are good grounds to accept that the moral rightness of an action is a reason to perform it, and the main grounds against accepting this view rest on a confused picture of moral metaphysics. So, with that confusion cleared up, let's go ahead and accept it.

This is a substantive position. The claim that rightness *cannot* be a reason for action has been used a premise in some important arguments. Notably, many people hold that an agent and/or her action can attain some positive evaluative status only if she acts "for the right reasons". If one holds such a view, and also assumes that rightness is not a reason for action – so, *a fortiori*, it cannot be among the *right* reasons – then it follows that when someone does the right thing because it's the right thing to do, she and/or her action do not attain the relevant positive status. Thus, buck-passing about rightness has been used to denigrate agents who are motivated by rightness *de dicto*. For instance, Julia Markovits (2010, p.207) offers an argument of this form for the claim that people who are motivated by rightness *de dicto* cannot perform acts with moral worth, and David Shoemaker (2007, p.88) offers an argument of this form for the claim that such people cannot be full-fledged members of the moral community. If an action's moral rightness *is* a reason to perform it, then these arguments are all unsound. That is my primary motivation for writing this paper. If the view that I have defended is correct, it shows that a family of criticisms of agents who do the right thing because it's the right thing to do are based on unsound arguments.

Of course, the reader may think that there is something else wrong with having and acting on explicitly moral motivations. But I take myself to have shown that, if there is

anything wrong with having and acting on explicitly moral motivations, it is not that agents who do so thereby fail to act for genuine normative reasons.

III: Accidentally Doing the Right Thing

1. A Tale of Two Finns

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 the Stormtrooper – 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*.[19] 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

[19] 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.

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.

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 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 properties of acts *F*, someone accidentally does an *F* thing if she has no idea that her act possesses property *F* when she performs it.
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 her act is morally right when she performs it. *(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.

One more complex contemporary gloss holds 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.

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.

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 historical and contemporary literatures. There may be no such thing. This would cast some doubt on the usefulness of these literatures.

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.[20]

[20] 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).

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 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; 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.[21] 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 with property *F* it is sufficient that (a) she is motivated to perform it by the fact that it has property *G* and (b) as a matter of metaphysical fact, the act's having property *G* makes it the case that it has property *F*. We are supposed to think that it is "no accident" or "no coincidence" that someone acts *F*-ly when, given this metaphysical relationship between the feature of the act that motivates her and the act's *F*-ness, it is no *surprise* that she acts *F*-ly.

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

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

[21] 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 encouraging me to clarify this point.

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 it has a further property (promise-keeping). Yet it still seems as though I *accidentally* perform an act with this further property. 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 keeping my promise, 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[22]: 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[23]: 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; she has accidentally lured the vampires to their death. (The author of this example names it "Accidental Slayer".)

[22] This example is adapted from Lackey (2008).
[23] This example is adapted from Riggs (2014).

SEMAPHORE DANCER[24]: 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 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 to perform an act by a feature that makes it the case that the act possesses another property. Yet it still seems natural to say that they *accidentally* perform acts with these further properties. 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*.[25]

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 with a

[24] This example is adapted from Carr (1979).

[25] I am not sure precisely which metaphysical relationship defenders of the new view take to obtain between right-making features and moral rightness. But, since the metaphysical relationships in these cases are slightly different, together they should cover all the bases.

certain property, though they are each motivated by the feature of their act that makes it the case that it possesses the relevant property?

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 has property *G*, and (b) as a matter of metaphysical fact, property *G* makes it the case that the act has property *F*, 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 does something *F*, though she is motivated to perform her act by its *F*-making feature. Indeed, I suspect that, the more emphasis we place on the fact that she has *no idea whatsoever* that the property motivating her is *F*-making, the more it will seem that she does an *F* 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 with property *F* 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 act possesses an accidental-seeming property 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. 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 one property of an act (protecting a person) makes it the case that the act has another property (rightness) 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 properties of acts *F*, someone accidentally does an *F* thing if she has no idea that her act possesses property *F* when she performs it.

We have seen that CLAIM holds even of cases in which an agent is motivated to perform an act by the very feature that makes it the case that the act is *F* (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 *F*-ness, she still accidentally does an *F* thing.

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 does grasp 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 their acts *lack* the relevant properties. But this would only make it seem clearer that they accidentally perform acts with the relevant properties. 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 intuition that there seems something praiseworthy about Huck Finn. That is true. 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 distinct 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 T_1 with being praiseworthy for performing an act of type T_2, where the act's being of type T_2 metaphysically constitutes its being of type T_1. I'll now explain what I mean by this.

When someone is praiseworthy for acting rightly, she is praiseworthy in virtue of a property of an act that she performs: its moral rightness. This is a normal sort of thing. We are often praiseworthy in virtue of the properties of acts that we perform. 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.[26]

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

[26] Thanks to Yongming Han for helpful discussion of this point.

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.[27] 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 φ-ing" is a success term; one can φ deliberately only if one does in fact φ. (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

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

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 iff 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 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 say that agents deserve a lot of blame for bad side-effects that are foreseen but unintended, but 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 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 do an *F* thing neither deliberately nor accidentally, if she foresees that her act has property *F*, but does not choose to perform it on the basis of its *F*-ness. 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 her act has a property that are better than having *no idea whatsoever* that it has the property. For example, she could have a vague inkling that her act has the property. Or she could have credence 0.2467 that it has the property. So there are a 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 her act has a certain property to count as *deliberately* performing an act with this property. 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 deny the knowledge condition, 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 doing something *F* requires only minimal foresight that an act has property *F*, if performing an act with property *F* is precisely what one was trying to do all along.[28] Analogously, someone's trying to act rightly and succeeding can amount to her deliberately doing the right thing even if she is not at all confident that her act is morally right.

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 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

[28] 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 her 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 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 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.

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 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 do something *F* without thinking furiously of the *F*-ness of one's activity throughout the duration of this activity; one's awareness of the *F*-ness can operate at the subpersonal level. On my view, so long as the agent's attitude toward her act's moral rightness is 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.

5. Coda: Counterexamples and replies

The best strategy open to defenders of the new view, in resisting my main argument, is to deny premise 2. They may do this by claiming that the principle is too strong and offering counterexamples. I am not in a position to anticipate and respond to all possible putative counterexamples. But I have now said enough to be in a position to survey three *families* of putative counterexamples, members of which are proposed to me quite frequently, and to explain my response to each of these families of putative counterexamples. That is what I will do in this final section.

First: The defender of the new view might suggest that *under-confident agents* provide a counterexample to the claim in my second premise. For example, an under-confident test taker may ace a test while having no idea that she is acing it – owing to her lack of confidence, she may believe only that she is doing moderately well. Similarly, an under-confident basketball player may throw a three-pointer without having any idea that he is doing so – owing to his lack of confidence, he may strongly doubt that his throw will reach the hoop, and may be amazed when it does.[29]

My response to counterexamples in this family is that, in fact, the agents do have some idea of what they are doing. Moreover, since acing the test and throwing the three-pointer is precisely what the agents were *trying* to do all along, they count as doing these things deliberately even though they had a low degree of confidence that their acts possessed the relevant properties. This is like the case of Finn deliberately saving Poe despite rationally having a low degree of confidence in his success, with the modification that, in cases involving under-confident agents, we stipulate that the agents' low degree of confidence is irrational. Of course, the test-taker and basketball player may not report that they are trying to ace the test and to throw a three-pointer; since they are under-confident, they may say that they are simply trying to do their best, or something along these lines. But these reports are somewhat disingenuous. When the basketball player sees his ball go through the hoop, he is not indifferent to this fact. He would not be equally happy throwing or not throwing the hoop just as long as he did his best. Rather, it is true that he is trying to do his best, but what he is hoping is that "his best" refers to throwing a three-pointer. That is what he really wants to do.

[29] Thanks to Maria Lasonen-Aarnio and Nathan Howard for suggesting these examples to me.

These cases are unlike the case of Huck Finn. Huck does not even believe that he *might* be acting rightly. It is not the case that acting rightly is precisely what he was trying to do all along, though he has a low degree of confidence that he will successfully act rightly by helping Jim to escape. If this were the case, Huck would be motivated by the (possible) rightness of his act, and the Kantian view would then straightforwardly entail that his act has moral worth. On the contrary, the literary interpretation favored by defenders of the new view emphasizes that Huck is *not* trying to act rightly. He is fully convinced that it would be morally right to turn Jim in, but he disregards this and helps Jim to escape anyway. There are versions of the test-taker and basketball player cases that are like this. For example, in school I deliberately under-performed in tests, so as to keep a low profile and thereby avoid bullying. I would figure out the correct answers, and then write other answers, or leave the question blank and draw pictures of penguins. If it had turned out that the answers that I thought were correct were incorrect, and that the answers that I wrote instead were correct (or that pictures of penguins were somehow correct), then I would have *accidentally* aced the test while trying to fail it. This is closer to Huck's case than the under-confident test taker, who is still trying to ace the test. But I am clearly not praiseworthy for acing the test in this case.

A second family of putative counterexamples to my premise center around *conceptual truths*. For example, suppose that somebody deliberately takes another person's bag without permission, and we accost her for stealing the bag. It seems odd for her to say, "Oh, I'm sorry! I knew that I was taking another person's bag without permission, but I didn't know that this constitutes stealing. So I guess I *accidentally* stole your bag!" Similarly, if a gardener deliberately prunes some hedges but does not possess the concept HORTICULTURE, it would be odd to say that he *accidentally* engaged in horticulture. Yet the thief professes to have no idea that her act is an instance of stealing, and the gardener,

lacking the concept HORTICULTURE, has no idea that his actions are instances of engaging in horticulture.[30]

My response to counterexamples in this family is to deny that the agents have no idea that their acts possess the relevant properties. They know full well that their acts possess the relevant properties. They simply lack or misunderstand some concepts by means of which we may refer to the relevant properties. But they are fully competent with *other* concepts by means of which we may refer to the properties. And, using those concepts, they predicate the properties of their acts: the thief cheerfully acknowledges that she is taking another person's bag without permission, and the gardener presumably knows that he is gardening. But it is a conceptual truth that stealing is taking another person's bag without permission, and it is a conceptual truth that horticulture is gardening. These are just different ways of referring to individual properties. So, these agents are still able to refer to the target properties and to predicate them of their actions. And this is precisely what they do. They would not recognize some descriptions of the properties as applying to their actions, but that is immaterial. Their confusion is conceptual, not metaphysical – it is about which terms refer to which properties, rather than about which properties their acts possess.

These cases are also unlike the case of Huck Finn. He does not predicate the property of moral rightness of his act, using other concepts than the concept MORALLY RIGHT. On the contrary, he does not recognize that his act instantiates this property at all. His confusion is metaphysical, not conceptual – it is about which properties his act possesses. We need to be careful here, because it is natural to say both that the thief knows that her act possesses a property (taking another person's bag without permission) that constitutes

[30] Thanks to John Schwenkler and Brendan Balcerak Jackson for suggesting these examples to me.

stealing, and that Huck knows that his act possesses a property (protecting a person) that makes it right. This might make it seem that the cases are on a par. But they are not. We must be careful with words like "constitutes", which are ambiguous between different metaphysical relationships. The property that the thief predicates of her action *is* the property of stealing; it is a conceptual matter that this is what stealing is. The descriptions "stealing" and "taking another person's stuff without permission" are just two terms for a single property. But the property that Huck predicates of his act is not the property of moral rightness. Rather, it is a lower-order property that bears a metaphysical relationship – the "makes it the case" relationship – to the fact that his act possesses the further property of moral rightness. And he has no idea about this metaphysical relationship. So, unlike the thief and the gardener, Huck has no idea that his act possesses the property of moral rightness under *any* description.

I acknowledge that it can be very difficult to distinguish between the metaphysical relationship that obtains when two concepts refer to a single property, and that which obtains when two concepts refer to distinct properties such that the lower-order one makes it the case that the higher-order one obtains. But there are ways of doing this. Here is how we can tell that it is not a *conceptual* truth that the property that motivates Huck makes it morally right: people who are fully competent with the concept of moral rightness disagree about whether this property is or is not a right-making feature. Their disagreement is substantive moral disagreement. It is not indicative of conceptual confusion. By contrast, the thief fails to recognize that her act is an instance of stealing only because of conceptual confusion, and the gardener fails to recognize that she is engaging in horticulture only because she does not possess the concept HORTICULTURE.

With the foregoing in mind, here is a test we can employ to determine when it is appropriate to say that someone has an idea that she is doing an *F* thing, using a

description of property *F* that she herself would not recognize as applying to her action (call this description "the target description"). First, take all the descriptions that the agent *does* recognizes as referring to properties of her action. Then, suppose that someone competent with all concepts in the target description was given all of this information about the agent's action. If the person competent with the concepts in the target description could infer, just from her conceptual knowledge plus the information that the agent recognizes, that the agent's action has property *F* considered under the target description, then it is appropriate to say that the agent does have an idea that she is doing an *F* thing, describing property *F* with the target description. If the agent's understanding of the nature of her act plus full competence with the concepts in the target description would be enough to recognize this description's application to her act, then the agent's ignorance is merely conceptual rather than metaphysical.

A third family of putative counterexamples, and the last one that I will discuss here, concerns descriptions of the *precise physical execution of skillful actions*. For example, return to the case of the basketballer. Perhaps, in throwing his three-pointer, the basketballer bends his knees at a certain precise angle – say, 68 degrees. He may well have no idea that he is bending his knees at this particular angle. But it seems odd to say that he *accidentally* bent his knees at this particular angle. Similarly, a virtuoso violinist may press down on the strings of her violin with a certain amount of force, but may never have given this fact a moment's thought. It seems natural to say that she has no idea that she is pressing down on the strings with this exact amount of force. But it seems strange to say that she does so accidentally.[31]

[31] Thanks to Jim Conant and Brian Weatherson for suggesting these examples to me.

My response to counterexamples in this family is to divide them in two. I think that a skillful agent really does have an idea of some aspects of the precise manner in which she physically executes her activity, and she does not act in the relevant ways accidentally. But I think that the skillful agent may really have no idea of some other aspects of the precise manner in which she physically executes her activity, and these really are accidental – indeed, I think it is unclear whether we should describe the agent as *doing* these things.

The previous point about different ways of referring to a single property can help here. When thinking of the physical execution of tasks that we perform skillfully, we may be able to refer to some aspects of the precise manner in which we perform this activity by simple ostension. The basketballer may think of the angle of his knees by thinking that he is bending his knees "*this* way". And the violinist may think of herself as pressing down on the strings "like *so*". In context, these descriptions refer to a 68-degree angle and a certain amount of force. Unfortunately, we cannot apply my test of conceptually competent speakers in these cases, since "this way" and "like so" are directly referring expressions which have no content independent of context. But I think it is nonetheless plausible that these skillful agents directly refer to certain facts about the precise physical execution of their actions by ostension. Moreover, since one need not think furiously of the fact that one is performing an act with a certain property in order to deliberately perform an act with this property, the basketballer and the violinist need not furiously ostend the physical manner in which they perform their actions in order to deliberately perform them in this manner. Their direct, ostend-able awareness of the precise physical manner in which they are performing their actions can operate at the subpersonal level.

Nonetheless, ostension is tricky, and it is frequently unclear what exactly an agent ostends. There may be aspects of the physical execution of her action on which the agent

has no grasp whatsoever. For instance, the violinist may have no grasp on physical facts about how the muscle fibers within her arms are moving in order to enable her to press down on the strings with a certain amount of force. At bottom, she certainly has no grasp of the microphysical facts about the subatomic particles that make up her arm and what they are doing. So it may be that she really has *no idea* that the muscle fibers within her arms are moving in a certain way, or that the subatomic particles have a certain speed and location; there is no description at all, not even one involving ostension, under which she would recognize that this is happening. In that case, I am inclined to say that the movement of her muscle fibers and the activity of the subatomic particles is accidental. I fully acknowledge that this still sounds a bit strange. But it is noteworthy that it sounds even more strange to say that she *deliberately* moves the muscle fibers in her arms or the subatomic particles constituting them, given that she has no awareness of any kind that this is occurring. And it is plainly false to say that she foresees that she is moving the muscle fibers or subatomic particles in the relevant way, but does not intend to do so. None of these descriptions sounds good. I think that this is because we should not even think of the movement of the muscle fibers in her arms or of the subatomic particles as something that the agent *does*. These are rather facts about how what she does is physically realized in the world.

The case of Huck Finn is not like these cases, either. Moral rightness is not a property pertaining to the precise physical details of the movements of his limbs whereby Huck helps Jim to escape. It is a property whose instantiation is made the case *by* the properties that Huck has in mind, rather than a property whose instantiation makes it the case that he performs an act with the properties that he has in mind. Huck moves his body in a certain physical way, thereby protects Jim, and thereby acts rightly. So, tricky questions about whether his moving his body in a certain way is accidental, deliberate, or foreseen but unintended do not tell us anything about whether it is accidental, deliberate, or

foreseen but unintended that he performs an act with the higher-order properties of protecting Jim and moral rightness. Whatever is the correct story to tell about an agent's relationship to the lower-order properties pertaining to the physical execution of that which does deliberately, there is no reason to expect this to be the same as the correct story to tell about an agent's relationship to the higher-order properties that are made the case by that which does deliberately. So I am not worried by this family of putative counterexamples.

Most of the putative counterexamples to my second premise that I have so far been offered fall into one of the three categories just discussed. This is, of course, not an exhaustive list of all putative counterexamples. But I hope that I have said enough to enable the reader to construct similar responses to other putative counterexamples on my behalf.

Here is where this leaves the dialectic between the Kantian and new views. The new view offers the case of Huck Finn as a case in which someone performs an act with moral worth but does not do the right thing because it is right, and thus a counterexample to the Kantian view. I have suggested that the case of Huck Finn is in fact a case of someone's merely accidentally doing the right thing, and thus not a case of an act with moral worth, and thus not a counterexample to the Kantian view at all. The defender of the new view will want to resist my claim that Huck accidentally does the right thing. Since they hold that he has no idea that his act is morally right, they must argue that it is not the case that someone accidentally does an *F* thing if she has no idea that her act has property *F* when she performs it. I have surveyed three ways to resist this claim, using three types of putative counterexample. I have argued that these families of counterexample all fail, and that none of them are very close to the case of Huck Finn anyway. It now falls to the defender of the new view to find some other way of resisting my main argument.

Concluding Remarks

In the preceding papers I argued that motivation by rightness *de dicto* is just as praiseworthy as motivation by rightness *de re*, that the fact that an act is morally right is a genuine objective normative reason to perform it, and that someone's trying to act rightly is necessary for her to count as deliberately doing the right thing, which is in turn necessary for her to be praiseworthy for doing the right thing and to perform acts with full moral worth. Together, these papers form part of the case in favor of the sort of explicitly moral motivation that I described in the introduction to this Dissertation. The full case will take several more papers, and will require exploring some new topics. Here is a summary of the directions in which I would like to take this project in future work.

One central question concerns what exactly an agent must have "in her head", to paraphrase the quotation from Arpaly discussed in the last paper, in order to count as being motivated to act rightly. I do not think that being motivated by a right-making feature is sufficient for someone to count as being motivated to act rightly; someone can be intrinsically motivated by a right-making feature while having no idea of, and being indifferent to, the fact that this feature is morally significant, so this is clearly not the sort of explicitly moral motivation that I have in mind. But I do hold that there are ways of being motivated to act rightly that do not involve the agent thinking of what she is doing under the concept MORALLY RIGHT. Moral rightness is a property, to which our attitudes may refer. Like other properties, it is possible to refer to moral rightness under some different descriptions. As I discuss in the first paper, I think it is plausible that the concept

of moral rightness is a cluster concept, and that there are a variety of conceptual connections between the concept of rightness and other normative concepts such that someone must grasp sufficiently many of these connections (which may be a vague matter) in order to count as referring to the property of moral rightness with her terms "rightness" and "right". The interesting questions concern how many and which of these connections one must grasp in order to count as being motivated to act rightly, and whether there is anything else someone could have in her head that would also count as being motivated to act rightly.

In this vein, one possibility that I intend to explore is the possibility that wanting to *strike the right balance* between multiple morally significant things that one sees are at stake, or to *respond appropriately* to the fact that they are all at stake and are important to different degrees, is a way of being motivated to act rightly. The property of moral rightness is the property that an act has when it strikes the right balance between all the morally significant things at stake in the agent's situation, constituting an appropriate response to the fact that these things are at stake and are important to different degrees. So, it seems plausible that wanting to strike the right balance or to respond appropriately to the many morally significant things at stake is a way of wanting to act rightly; it is something that can be in an agent's head that is sufficient for her to count as wanting to act rightly. I find this promising because I think that it is an empirically observable fact that ordinary people who face morally charged situations often respond by deliberating about which of the things at stake is most important, what relationships these things bear to one another, and so on. We do not typically respond to moral conflict by shrugging our shoulders and using a coin-flip or some other randomizing device in order to decide what to do. I think that this is because, in addition to caring about the particular morally significant things at stake, we care about striking the right balance between them. Given the conceptual connections just described, this is itself a species of trying to act rightly. It

is not quite motivation by rightness *de dicto*, since the agent's motivation need not include the concept MORALLY RIGHT. But it is a motivation that has as its object the property that in fact is the property of moral rightness. I am interested in this possibility because it would show that the kind of moral motivation that I defend is much more widespread, and much less odd, than some of its detractors suggest.

In this connection am also interested in the possibility that there is a characteristic phenomenology associated with moral motivation. This idea is offered in a much more tentative spirit. My sense is that there is a certain character to the experience of seeing a consideration as calling for a certain response *morally* – seeing it as a *moral* reason – which unifies moral experience but differs from the character of the experience of seeing a consideration as calling for a response on, say, prudential or epistemic or aesthetic grounds. I suspect that this characteristic experience, to the extent that it is unified across cases of recognizing a certain course of action as called for on moral grounds, helps young children to initially grasp the concept of moral rightness. If this is correct, then it suggests a very different kind of way in which someone might count as motivated to act rightly: she might come to recall and to recognize the distinctive character of moral experience, after repeated experiences of this sort, and then think that she wants to perform actions like *that*. This might be a way to directly refer to the property of moral rightness. If that is correct, then someone may count as being motivated to act rightly with only a very minimal degree of cognitive and conceptual sophistication. I am not sure whether this idea works, but it is one that I am interested in exploring.

Another set of questions concerns the distinctions I have drawn between the many ways in which an agent can be morally praiseworthy or blameworthy, and the ways in which they go together or come apart. I am particularly interested in further exploring the question, raised in the first paper, of what to say about agents who are well-meaning but

morally mistaken – who are motivated by a moral feature *de dicto* but not *de re*, as they are mistaken about what this feature consists in. I noted there that we may say that someone's motivation is still praiseworthy even if her wrongful action and her false moral belief are blameworthy. But I did not address the question of when moral ignorance is itself blameworthy. This is a huge question, about which there is a huge literature. The contemporary literature is divided between extreme positions: there is a position informed by the voluntarist approach to thinking about moral responsibility, according to which people are *almost never* blameworthy for moral ignorance, and there is a position informed by the quality-of-will approach to thinking about moral responsibility, according to which people are *always* blameworthy for moral ignorance. I think that this second position rests on a misunderstanding of the quality-of-will view's implications, among both its detractors and its proponents. The idea is supposed to be that moral ignorance is always blameworthy because it manifests a failure to care adequately about what is morally significant. But, I maintain, moral ignorance does *not* always manifest a failure to care adequately about what is morally significant. Sometimes it does reflect such a failure – motivated ignorance is sadly widespread. But there are also lots of real-life cases in which the delicacy and complexity of the moral truth ensure that even someone who cares far more than "adequately" about all morally significant features of her situation can be mistaken about what response they call for. In these cases, I maintain, the agent's moral ignorance is not blameworthy, as it does not manifest a failure to care adequately. Reflection on these cases can also help us to think about what it is to care adequately. I will argue for all of this in a future paper.

The third paper of my dissertation distinguished two ways for an agent to be praiseworthy: someone can be praiseworthy for deliberately performing an action of a good type, or she can be praiseworthy for having a character trait. I would like to explore this distinction in more detail in future work. I think that it captures a sense in which both

voluntarist and quality-of-will approaches to thinking about moral responsibility are on to something. In short, I think that we deserve moral credit *both* for what we deliberately do or cultivate *and* for the quality of our will. I think that these are genuinely two independent ways of being praiseworthy, neither of which is reducible to the other. But they are interrelated in important ways. First, the quality of our will is rarely something that is innate or wholly the product of cultural conditioning; it often is itself something that we have deliberately cultivated. So, while someone is directly praiseworthy for her good will, she may *also* be praiseworthy for having successfully cultivated it. Second, someone's deliberately performing an act of a certain good type typically requires her to have a certain quality of will. If the argument of my third paper is correct, one can deliberately perform an act of a good type only if one was *trying* to perform an act of this type, which requires *wanting* to perform an act of this type. And wanting to perform a good type of act is itself a praiseworthy motivation – a form of good will. So, while someone is directly praiseworthy for having deliberately performed a good type of act, the fact that she was trying to perform this type of act (*de dicto*) may also be a character trait that is praiseworthy in and of itself. I intend to further explore these two types of praiseworthiness and the relationships between them in future work.

I also intend to further defend the idea that it is morally praiseworthy to *try* to perform acts with positive moral features, whether or not one succeeds (though it may be more praiseworthy overall to try and succeed than to try and fail). In this vein I intend to write a response to a recent paper by Berislav Marušić arguing that thinking of oneself merely as *trying* to φ, rather than thinking of oneself as φ-ing, can exhibit a kind of bad faith. Marušić argues that merely promising to try to φ rather than to φ, when it is "entirely up to us" to φ, is wrong because it "hides a possible choice under the veil of our susceptibility to circumstances beyond our control" (2017, p.249). I agree with this as a claim about cases in which it is *entirely up to us* whether to φ; that is to say, in which we know exactly

what it would take to φ and nothing prevents us from taking these steps. But I think that few real-life cases are like this. And I think that it exhibits a praiseworthy kind of humility and honesty about one's chances of success to merely promise to try to φ, and to think of oneself as trying to φ rather than as φ-ing, when there are *salient obstacles* that may prevent one from φ-ing. This is particularly so, I think, when one faces *salient epistemic obstacles*: when it is unclear whether one knows or will know exactly what it would take to φ. For instance, it is appropriate for parents to ask their child to promise to try their hardest to spell all the words correctly in a difficult spelling bee, but it would be weird (and inappropriately sensitive to the salient epistemic obstacles) for them to ask the child to promise to spell all the words correctly. I also think that, when it comes to acting morally rightly – or to acting fairly, or kindly, or to promoting well-being, or etc. – there are always salient epistemic obstacles. It is always unclear whether we know what it takes to perform acts with any of these positive moral features. Hence, in the moral case, it is always appropriate to think of ourselves as trying.

Most of the paper topics just discussed are in some way related to the phenomenon of moral uncertainty: wanting to perform acts with some positive moral feature, but being uncertain as to precisely what it is for acts to have this feature, and thus being uncertain which acts have this feature. I am very interested in the question of what exactly trying to act rightly amounts to for agents with a high degree of moral uncertainty. If someone thinks that she knows what the right thing to do is, or at least thinks that she has a pretty good guess, then she can choose to do it *because it's the right thing to do*. But if someone is too morally uncertain to see any of her options as a good moral guess, and yet she wants to act rightly, then it is unclear what she should do. In this vein I am particularly interested in the debate between internalist and externalist ways of thinking about moral uncertainty. Very roughly, internalist responses are those that see an agent's moral uncertainty as itself morally relevant, such that what morally uncertain agents should do

differs from what morally certain agents should do, whereas externalist responses deny this. There are a variety of internalist positions, each corresponding to a different decision-rule that takes account of an agent's credences in moral theories when identifying a choice-set from among her available actions. I am sympathetic to the internalist approach. I am particularly interested in a curious problem for internalists: the problem lies in adapting internalist decision-rules to take account of a kind of *meta-level* uncertainty about whether the internalist approach is, indeed, correct. It turns out to be quite difficult for internalists to satisfactorily take account of the possibility that externalism is true. I explore this problem in another paper.

Lastly, I am interested in further exploring the picture of moral metaphysics that I sketch in the first paper and spell out in more detail in the second paper: the view that I call the "share the weight" view. There are a lot of unanswered questions that a full defense of this view would have to answer. One concerns how to draw the boundaries around facts in a metaphysical hierarchy to distinguish those that share a particular "chunk" of normative weight from those that do not. Intuitively, not *all* of the facts in any given hierarchy are reasons: some are too low-down, so to speak, while others are too high-up. For instance, facts about the speed and location of fundamental physical particles may appear at the very bottom of all metaphysical hierarchies (if moral naturalism is true), but intuitively these facts are not reasons. Similarly, the fact that an act is morally right makes true all possible disjunctions with the fact that the act is morally right as one disjunct: the fact that the act is morally right *or* fluffy, the fact that the act is morally right *or* I am wearing a hat, and so on. But, intuitively, these facts are not reasons. I think that, properly understood, the task of drawing the boundaries around chunks of normative weight just is the task of doing first-order normative ethics: in figuring out which things count morally in favor of which responses, and which of them bear metaphysical relationships to one another such that their normative weight is shared, we are

identifying the first-order normative facts. But this does not mean that there are no useful general tests we can employ to see where the boundaries lie. I think that there are two tests, each of which I briefly mention in the second paper. We can imagine an agent motivated to perform an act by a certain fact, and can ask ourselves whether this elicits the response characteristic of mismatch between an agent's motivating and normative reasons. We also can imagine that someone learns one fact about an act and nothing else, and can ask ourselves whether this intuitively alters her normative situation (as in the example of the big red button). Again, this is something that I would like to explore further in future work.

Another significant unanswered question for the "share the weight" view concerns what to say about the difference between moral reasons and other types of reasons. I have argued that an act's moral rightness is a reason to perform it. Presumably, it is a *moral* reason. Nonetheless, an act's moral rightness may *also* be a reason of another kind. For instance, if I am going to be paid $100 for each morally right act that I perform, then an act's rightness may be a prudential reason to perform it. And if I find the performance of morally right acts stunningly beautiful, then the fact that an act is morally right may be an aesthetic reason to perform it. It is unclear what exactly this difference between different types of reasons amounts to. As I see it, there are two main possibilities. One is that there are different types of normative weight, each corresponding to a different type of reason: moral weight, prudential weight, aesthetic weight, and so on. To the extent that there are differences between experiencing a consideration as calling *morally* for a certain response and experiencing it as calling for a response in these other ways, these differences in the character of our normative experience might support the idea that there are different types of normative weight. The second possibility is that there is just one type of normative weight, and the differences between different types of reasons stem from differences in the nature of the facts that they make the case – that are, so to speak,

the "top fact" in a hierarchy sharing a particular chunk of normative weight. On this second possibility, moral reasons are those that occur in hierarchies whose top fact is a moral fact, whereas epistemic reasons are those that occur in hierarchies whose top fact is an epistemic fact, and so on. To the extent that we are able to commensurate and compare the strength of different types of normative reason, this commensurability might support the idea that there is just one type of normative weight. I am not yet sure which of these ideas I prefer. Figuring this out, then, is another task for future work.

In sum, there is clearly a great deal more to be done to develop the research program that begins with this Dissertation. Nonetheless, I hope that my work so far has already gone some way toward rehabilitating the status of explicitly moral motivation in the eyes of contemporary ethicists and metaethicists. I look forward to praising explicit moral effort and discussing the nature of moral achievement for many years to come.

References

Aboodi, Ron (2016). "The Wrong Time to Aim at What's Right: When is *De Dicto* Moral Motivation Less Virtuous?" *Proceedings of the Aristotelian Society* 115(3), 307-314.

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.

Bedke, Matthew (2011). "Passing the Deontic Buck". *Oxford Studies in Metaethics* 6, 128-151.

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

Bennett, Jonathan (1988). *Events and Their Names*. Indianapolis: Hackett.

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

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

Carbonell, Vanessa (2013). "De Dicto Desires and Morality as Fetish". *Philosophical Studies* 163, 459-477.

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

Copp, David (1997). "Belief, Reason, and Motivation: Michael Smith's *The Moral Problem*". *Ethics*, 108(1), 33-54.

Crisp, Roger (2005). "Value, reasons and the structure of justification: How to avoid passing the buck." *Analysis* 65, 80-85.

Cuneo, Terence and Shafer-Landau, Russ (2014). "The Moral Fixed Points: New Directions for Moral Nonnaturalism." *Philosophical Studies* 171(3), 399-443.

Dancy, Jonathan (2000). "Should we pass the buck?" In A. O'Hear (ed.), *Philosophy: The good, the true, and the beautiful*. Cambridge: Cambridge University Press.

Darwall, Stephen (2010). "But It Would Be Wrong." *Social Philosophy and Policy* 27(2), 135-157.

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.

Dreier, James (2000). "Dispositions and Fetishes: Externalist Models of Moral Motivation". *Philosophy and Phenomenological Research*, 60(3), 619-638.

Enoch, David (2011). *Taking Morality Seriously*. Oxford: Oxford University Press.

Fogal, Daniel (2017). "Reasons, Reason, and Context". In E. Lord & B. Maguire (Eds.), *Weighing reasons*. Oxford: Oxford University Press.

Fried, Charles (1970). *An Anatomy of Values.* Cambridge, MA: Harvard University Press.

Gibbard, Allan (1990). *Wise Choices, Apt Feelings*. Cambridge, MA: Harvard University Press.

Gregory, Alex (2014). "A very good reason to reject the buck-passing account." *Australasian Journal of Philosophy* 92, 287-303.

Harman, Elizabeth (2011). "Does Moral Ignorance Exculpate?" *Ratio* 24, 443-468.

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.

Hills, Alison (2009). "Moral Testimony and Moral Epistemology". *Ethics* 120, 94-127.

Horgan, Terence and Mark Timmons (1990). "New Wave Moral Realism Meets Moral Twin Earth". *Journal of Philosophical Research,* 16, 447-465.

Horgan, Terence and Mark Timmons (1992). "Trouble on Moral Twin Earth: Moral Queerness Revisited". *Synthese*, 92(2), 221-260.

Horty, John F. (2012). *Reasons as Defaults*. Oxford: Oxford University Press.

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

Keshet, Ezra and Florin Schwarz (*ms*). *De re* / *De dicto.* Unpublished manuscript.

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.

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

Korsgaard, Christine (1983). "Two Distinctions in Goodness". *The Philosophical Review* 92(2), 169-195.

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

Lewis, David (2000). "Causation as Influence". *The Journal of Philosophy* 97(4), 182-197.

Liao, S. M. (2009). "The buck-passing account of value: lessons from Crisp." *Philosophical Studies* 151, 421-432.

Lillehammer, Hallvard (1996). "Smith on Moral Fetishism". *Analysis*, 57(3), 187-195.

Quine (1956). "Quantifiers and Propositional Attitudes." *Journal of Philosophy* 53(5), 177-187.

Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong*. Harmondsworth: Penguin.

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

Marušić, Berislav (2017). "What's Wrong with Promising to Try?" *Pacific Philosophical Quarterly* 98(SI), 249-256.

McDowell, John (1998). "Non-Cognitivism and Rule-Following". In *Mind, Value, and Reality.* Cambridge, MA: Harvard University Press.

Mill, John Stuart (1871). *Utilitarianism*, 4th ed. London: Longmans, Green, Reader, and Dyer.

Nair, Shyam. (2016). "How do reasons accrue?" In E. Lord & B. Maguire (Eds.), *Weighing reasons*. Oxford: Oxford University Press.

Olson, Jonas (2002). "Are Desires *De Dicto* Fetishistic?" *Inquiry* 45(1), 89-96.

Olson, Jonas (2004). "Buck-passing and the wrong kind of reasons." *Philosophical Quarterly* 54, 295-300.

Putnam, Hilary (2002). "The Entanglement of Fact and Value". In *The Collapse of the Fact/Value Dichotomy and Other Essays*. Harvard, MA: Harvard University Press.

Parfit, Derek (2001). "Rationality and reasons." In D. Egonnson, J. Josefsson, B. Petersson and T. Rønnow-
Rasmussen (eds.), *Exploring Practical Philosophy: From Action to Values*. Aldershot: Ashgate.

Rabinowicz, Wlodek, and Ronnow-Rasmussen, Toni (2000). "A Distinction In Value: Intrinsic and For its Own Sake". *Proceedings of the Aristotelian Society* 100(1), 33-51.

Rabinowicz, Wlodek and Rønnow-Rasmussen, Toni. (2004). "The strike of the demon: on fitting pro-attitudes and value." *Ethics* 114, 391-423.

Railton, Peter (1981). "Probability, Explanation, and Information". *Synthese* 48, 233-256.

Railton, Peter (1986). "Moral Realism". *The Philosophical Review* 95(2), 163-207.

Railton, Peter (1993) "Noncognitivism about Rationality: Benefits, Costs, and an Alternative." *Philosophical Issues* 4, 36-51.

Rawls, John (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Raz, Joseph (1990). *Practical Reason and Norms.* Princeton, NJ: Princeton University Press.

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

Roberts, Debbie (2013). "Thick Concepts". *Philosophical Compass* 8, 677-688.

Scanlon, Thomas (1998). *What We Owe To Each Other.* Cambridge, MA: Harvard University Press.

Shoemaker, David (2007). "Moral Address, Moral Responsibility, and the Boundaries of the Moral Community." *Ethics* 118(1), 70-108.

Schroeder, Mark (2009). "Buck-passers' negative thesis". *Philosophical Explorations* 12, 341-347.

Skorupski, John (2007). "Buck-passing about goodness." In T. Rønnow-Rasmussen, J. Josefsson, D. Egonsson, and B. Petersson (eds)., *Hommage à Wlodek; Philosophical papers dedicated to Wlodek Rabinowicz*. Published as web resource. URL: -

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.

Smith, Michael (1996). "The Argument for Internalism: Reply to Miller". *Analysis,* 56, 175-184.

Strandberg, Caj (2007). "Externalism and the Content of Moral Motivation". *Philosophia* 35, 249-260.

Svavarsdóttir, Sigrún (1999). "Moral Cognitivism and Motivation". *Philosophical Review* 108, 161-219.

Swanson, Eric (2010). "Lessons From the Context Sensitivity of Causal Talk". *The Journal of Philosophy* 107(5), 221-242.

Suikkanen, Jussi (2004). "Reasons and value—in defence of the buck-passing account." *Ethical Theory and Moral Practice* 7, 513-535.

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

Stratton-Lake, Philip (2002). "Introduction." In P. Stratton-Lake (ed.), *Ethical Intuitionism: Re-evaluations*, pp.1-28. Oxford: Clarendon Press.

Stratton-Lake, Philip (2003). "Scanlon's contractualism and the redundancy objection." *Analysis* 63, 70-76.

Stratton-Lake, Philip and Hooker, Brad (2006). "Scanlon versus Moore on goodness." In T. Horgan & M. Timmons (eds.), *Metaethics after Moore*, pp. 149-168. Oxford: Clarendon Press.

Toppinen, Teemu (2004). "Moral Fetishism Revisited." *Proceedings of the Aristotelian Society* 104(3), 305-313.

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

Väyrynen, Pekka (2006). "Resisting the buck-passing account of value." *Oxford Studies in Metaethics* 1, 296-324.

Väyrynen, Pekka (2013). *The Lewd, The Rude and The Nasty: A Study of Thick Concepts in Ethics*. Oxford: Oxford University Press.

Weatherson, Brian (2014). "Running Risks Morally". *Philosophical Studies* 167(1), 141-163.

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

Yablo, Stephen (1992). "Mental Causation". *The Philosophical Review* 101(2), 245-280.

Zangwill, Nick (2003). "Externalist Moral Motivation". *American Philosophical Quarterly*, 40(2), 143-154.