Asian Journal of Philosophy (2025) 4:11
https://doi.org/10.1007/s44204-024-00230-2

ARTICLE SYMPOSIUM

A theory of assessability for reasonableness

Andrew T. Forcehimes[1]

Received: 9 November 2022 / Accepted: 3 December 2024

Abstract
This essay defends an account of what things are assessable for reasonableness and why. On this account, something is assessable for reasonableness if and only if and because it is the functional effect of critical reasoning.

Keywords Normativity · Reasonableness · Reasoning · Functions

1 Introduction

Reasonableness lies at the heart of normativity.[1] To claim that you are being reasonable is to claim that you have lived up to the standards of reason. To claim that you are being unreasonable is to claim that you haven't. Yet not everything is open to assessments of reasonableness. The actions and attitudes of cognitively unsophisticated animals are neither reasonable nor unreasonable. They are areasonable, falling outside the domain of reason.

This essay is concerned with locating and explaining this boundary. What features must a thing have to be assessable for reasonableness? That is, what does it take for a thing to possess the property of being reasonable or being unreasonable (as opposed to being areasonable)? Call this the *possession question*. This question gives rise to another. For whatever features we identify in our answer to the possession question, the connection between reasonableness and these features is the sort of connection that calls for an explanation. We thus need to explain why the identified features make a thing assessable for reasonableness. Call this the *explanatory question*. If we could answer both the possession question and the explanatory question —giving an account of what things are assessable for reasonableness and why—we would have a *theory of assessability for reasonableness*.

[1] Here I am following Scanlon (1998) in normative ethics and Chisholm (1989) in epistemology.

✉ Andrew T. Forcehimes
forcehimes@ntu.edu.sg

[1] Nanyang Technological University, 48 Nanyang Avenue, HSS 03-31, Singapore 639818, Singapore

Published online: 07 January 2025

A theory of assessability for reasonableness gives us a principled way of sorting those things that are capable of being assessed for reasonableness (the reasonable and the unreasonable) from those that are not (the areasonable). Such a theory is importantly different from a substantive theory of what makes something reasonable or unreasonable. In normative ethics, a theory telling us that, of the actions assessable for reasonableness, actions that treat persons as mere means are unreasonable is a substantive theory. In epistemology, a theory telling us that, of the beliefs assessable for reasonableness, only those supported by the agent's evidence are reasonable is a substantive theory. As these examples suggest, a theory of assessability for reasonableness is prior to substantive theories in the sense that they identify their range of application. A theory of assessability for reasonableness tells us which things fall within the domain of reason, and so which things a substantive theory, to be complete, needs to assess.

Below I begin by motivating a handful of claims that will help us home in on an answer to the possession question (§2). These claims do not take us all the way to the view that I ultimately hope to defend. But they eliminate rivals. And they are suggestive of the view that assessability for reasonableness demands a connection to reasoning. I then turn to whether this view can give us a satisfying answer to the explanatory question (§3). Appealing to the function of reasoning, I argue that it can.

The theory I plan to defend can be described roughly in the following way: The answer to the possession question is that something is assessable for reasonableness if and only if it is caused in the right way by reasoning. And the answer to the explanatory question is that something is assessable for reasonableness if and only if it is caused in the right way by reasoning, because for it to be caused in the right way by reasoning just is for it to be caused by reasoning performing its function by way of the mechanism that was used when it was selected for.[2] And when something is caused by reasoning performing its function by way of the mechanism that was used when it was selected for—when something is the functional effect of reasoning—that thing becomes subject to the standards relevant to that function. It becomes subject to the standards of reason.

Before turning to the argument, a preliminary remark is in order. Attitudes and actions are the types of things we usually assess for reasonableness. But not all attitudes and actions are assessable. This has important implications for how best to proceed. It makes attitudes and actions the place to focus. It is not informative to compare the assessability of rocks to the attitudes of mature human beings. We won't be able to pinpoint precisely what makes the one and not the other assessable. Their differences are too extreme. I focus on attitudes and actions with the hope that we will be able to locate exactly what makes some, but not others, assessable for reasonableness. In the end, however, we might come to discover that things other than attitudes and actions—e.g., feelings—are assessable for reasonableness. Or perhaps, we will discover that all attitudes or all actions are areasonable. I am open to these possibilities. I concentrate on attitudes and actions only because they seem the most fruitful place to begin our inquiry.

[2] For the sense of "selected for" I have in mind, see Neander (2009) and McLaughlin (2007).

2 The possession question

In answering the possession question, the first claim I aim to defend is

> *Causal sequence exclusivity*: Whether something is assessable for reasonableness depends exclusively on the actual sequence causing or causally sustaining it.[3]

Some initial plausibility for causal sequence exclusivity comes from the information we usually rely on in making judgments about assessability. Often knowing a thing's causal origins is sufficient to make such a judgment. Your friend washes her hands because she has a brain tumor. This we judge as areasonable because of its etiology: it was caused by a tumor. You believe that washing your hands will help prevent the spread of disease after careful reasoning. We judge this belief as assessable for reasonableness because it was the upshot of reasoning.

Further support for causal sequence exclusivity comes from Frankfurt Cases. Consider

> **Frankfurt safety system**: You are reasoning about whether to wear your seatbelt. You recall videos of crashes where those without seatbelts flew through the windshield, you look at the warning signs on your car, you think about objects in motion staying in motion, and so on. Based on these considerations, you conclude that a seat belt would help in a crash. You put on your seatbelt. Unbeknownst to you, your car's safety system scans your brain. It monitors for a neurological pattern that you invariably exhibit prior to believing that a seat belt would help in a crash and putting on your seatbelt. If the pattern is not detected, then the system will initiate a neurological intervention, guaranteeing that you so believe and put on your seatbelt. As it happens, you believe and act on your own. You thus exhibit the pattern and the safety system remains dormant.[4]

Does the safety system make a difference to assessability? The answer seems to be No.[5] Your belief is assessable for reasonableness. Your putting on your seatbelt is assessable for reasonableness. The presence of the safety system made no causal difference to your holding the attitude or performing the action.[6] And because the

[3] This principle is modified from Sartorio (2017, 152) and Fischer (2012, 10).

[4] This case is modified from Widerker and McKenna (2006, 4).

[5] One forceful objection comes from the thought that answering in this way commits us to denying ought-implies-can. This objection is pressed by Widerker (1991), Copp (1997), and Haji (2002). Admittedly, the claim that an action cannot be wrong unless the agent could have acted differently has some ring of intuitive plausibility to it. Deontic notions govern sets of alternatives open to the agent. When an act is wrong, it is wrong because the agent ought to have acted in some other way. Wrongness, we might say, is mistake-implying. And some act cannot be a mistake unless there was some other alternative that was not a mistake. For, without the possibility to avoid it, something cannot be a mistake. Hence, when an agent is wrong this implies that the agent could have acted differently. For an especially clear articulation of this line of thought, see Haji (2023, §1.1). However, I think this objection can be resisted by appealing to Frankfurt-style omissions cases in the way described by Fischer (2003). Moreover, unlike deontic notions, being reasonable or being unreasonable does not carry the suggestion of governing a set of alternatives. In this respect, reasonableness is closer to intelligence. Being unintelligent or unreasonable is not mistake-implying. This is why the claim that action cannot be unreasonable unless the agent could have acted differently does not have the same ring of intuitive plausibility.

[6] For insightful discussion of this line of thought, see McKenna (2008, 771–772; 778–780).

safety system made no difference to the causal sequence leading to the formation of the attitude or the performance of the action, it made no difference to assessability.[7] As Frankfurt writes, "Why should [a] fact be considered in reaching a moral judgment concerning the person when it does not help in any way to understand […] what made him act as he did?" (1969, 837). If an assessment of reasonableness is warranted when the safety system is absent, that same assessment is warranted when the safety system is present. Assessability hinges on features of the causal chain alone.

Causal sequence exclusivity rules out any answer to the possession question that makes assessability depend on factors not contained within the actual causal sequence leading to the holding of the attitude or the performance of the action. Consider, for example, the view that an attitude or action is assessable for reasonableness if and only if (i) if the agent were to judge there to be sufficient reason to hold it or perform it, then she would hold it or perform it and (ii) if she were to judge there to be sufficient reason not to hold it or perform it, then she would not hold it or perform it.[8] On this judgment-sensitivity view, your belief that a seat belt would help in a crash is assessable if you believe because you judged there to be a sufficient reason to so believe; and if you had judged there to be sufficient reason not to so believe, then you would not have believed. On the face of it, this is a satisfactory verdict.

But return to Frankfurt safety system. Here your attitude and action, because of the presence of the safety system, are not judgment-sensitive. Had you judged there to be sufficient reason not to hold the belief, you still would have ended up holding it. For, if you didn't exhibit the neurological pattern, the safety system would have been activated, intervening to ensure that you believe that a seatbelt would help in a crash. (This is not what in fact happened. In the actual case, you were in no way interfered with. The safety system sat idly by. The movements of your mind were all your own.) So the judgment-sensitive view gives us the wrong results. It yields the mistaken verdict that your belief and action are areasonable. And we can identify where the view goes wrong. Because it makes assessability for reasonableness depend on how the agent responds counterfactually, it violates causal sequence exclusivity.

Causal sequence exclusivity points us in a general direction. If what it takes to be assessable for reasonableness depends on having a certain causal history, then, if we go back far enough (e.g., the Proterozoic Eon), we will find a time where everything is areasonable. The same holds locally for any given agent. If we go back far enough in the agent's history (e.g., infancy), we will find a time where nothing is accessible for reasonableness. So, assuming that a least some things are now assessable, there must be some part of the causal chain which converts the areasonable to things assessable for reasonableness. To answer the possession question, we need to locate precisely this part. I now want to suggest that we can exclude parts of the causal chain that admit of degrees with no non-arbitrary cut-off.

Assessability for reasonableness is all or nothing. The holding of an attitude, for example, may gradually slide from being unreasonable to being reasonable. But the holding of an attitude cannot gradually slide from being unreasonable to being

[7] For additional elaboration and defense, see Sartorio (2016, §3.1-§3.2).

[8] This view is inspired by Scanlon (1998, 2007).

areasonable. Being assessable for reasonableness is not like being red rather than orange. It is like being three-dimensional rather than two-dimensional. Is this attitude or action apt for reasonableness? This question calls for a definite yes or no answer: The answer can't be indeterminate or a matter of degree. To claim that an attitude or action is partially, sort of, somewhat, or a bit assessable for reasonableness shows that one fails to properly understand the notion.

Given that assessability for reasonableness has a precise boundary, we should accept the

> *Binary requirement:* Whether something is assessable for reasonableness depends on a factor that is either (i) all or nothing or (ii) admits of degrees but has a non-arbitrary cut-off.

Since assessability for reasonableness is all or nothing, we need a feature that can undergird this sharp break. This is because an answer to the possession question that violates the binary requirement would make the explanatory question unanswerable. If assessability is tied to a feature that admits of degrees with only arbitrary cut-offs, then when we go to explain why this precise degree marks the beginning of the domain of reason, our explanation would equally apply to a greater or lesser degree of that feature. And this would mean we have, in point of fact, failed to offer an explanation at all. Arbitrariness defeats explanation.

In this respect, it is helpful to compare possessing reasonableness with having freedom (or control). It is plausible to hold that whether an agent's actions or attitudes are free (or under her control) depends on something that admits of degrees. This is plausible because the notion of having or lacking freedom (or control) itself admits of degrees. Having freedom (or control) can fade into non-possession.[9] That is, you can slide from possessing complete freedom (or complete control) to completely lacking freedom (or being completely out of control). Consider the levels of freedom (or control) you possessed over your attitudes or actions as you grew from child to adult. A natural thought to have as we move along this continuum is that one becomes more and more free (more and more in control) as one develops. This is a natural thought because our notion of possessing freedom (or control) itself admits of degrees. The question "Is this person free (or in control)?" does not call for a definite answer. By contrast, it is baffling to claim that as a child matures a given attitude of hers becomes more and more assessable for reasonableness until eventually it is fully assessable. And if she later suffers from a degenerative brain disease, the levels of freedom (or control) she possesses may fade into non-possession, but it betrays a deep mistake to hold that as the disease progresses a given attitude becomes only partly assessable for reasonableness and eventually barely assessable at all.[10]

[9] For a compelling defense of this claim, see Mele (2006, 129–133). I borrow the example of developing from child to adult from him.

[10] In developing a theory of assessability for reasonableness, it is tempting to port over an existing account of freedom (or control) from the moral responsibility literature. But the difference described above—that freedom (or control) admits of degrees while assessability for reasonableness does not—should make us skeptical that such a strategy will succeed.

This requirement puts pressure on several views. Consider, for example, the view that an action or attitude is assessable for reasonableness if and only if it is caused by a mechanism that is moderately reasons-responsive, where to be moderately reasons-responsive is to be regularly reasons-receptive and weakly reasons-reactive. A mechanism is regularly reasons-receptive if and only if there is some suitably wide set of nomologically identical worlds where there is sufficient reason for the agent to hold the attitude or perform the action and she recognizes and assesses, through this mechanism, that there is sufficient reason to hold it or perform it. The mechanism is weakly reasons-reactive if and only if there are some suitable subset of the members of these worlds where the agent, through the mechanism, comes to hold the attitude or perform the action.[11]

On this reasons-responsiveness view, to be assessable for reasonableness, the causally operative mechanism must be moderately reasons-responsive. A mechanism being moderately reasons-responsive depends on how the mechanism behaves. If it behaves in certain ways in a suitably wide range of nomologically identical worlds, it will qualify as moderately reasons-responsive. But an account that determines whether something can possess reasonableness by appeal to how a mechanism behaves in a suitably wide range of nomologically identical worlds is an account that looks posed to run afoul of the binary requirement. For the continuum from completely reasons-unresponsive to completely reasons-responsive is smooth. There thus seems little hope of locating, in a non-arbitrary way, the precise threshold for where assessability for reasonableness begins. Wherever the purported threshold is located, we can always ask whether the range of suitable worlds cannot be slightly expanded or slightly reduced. The difference between what lies on either side of the suitability line is negligible, and such a negligible difference cannot ground being in or out of the domain of reason.[12]

Accordingly, if we adopted the reasons-responsiveness view as our answer to the possession question, we would render ourselves incapable of giving an adequate answer to the explanatory question. Suppose we claimed that $< W_1 \ldots W_n >$ constitutes the suitable set of worlds for assessability. How would we explain why exactly this set marks where assessability begins? The differences between $< W_1 \ldots W_n >$ and $< W_1 \ldots W_{n+1} >$ or $< W_1 \ldots W_{n-1} >$ are insignificant, so whatever explanation is offered for $< W_1 \ldots W_n >$ would equally apply to $< W_1 \ldots W_{n+1} >$ or $< W_1 \ldots W_{n-1} >$. As far as providing an explanation is concerned, the arbitrariness of the chosen set renders us impotent.

We are looking for a binary part of the causal sequence that makes something assessable for reasonableness. The obvious place to look is the agent's psychology.

[11] This view is modified from Fischer and Ravizza (1998). For attempts to extend this view to attitudes, see McCormick (2014) and McHugh (2017).

[12] The binary requirement also poses problems for the view defended by Smith (2003), Wedgwood (2017, §3.3), and Portmore (2019, 96). On this rational capacities view, something is assessable for reasonableness if and only if the agent has certain rational capacities such that "under a suitably wide range of counterfactual conditions, she would recognize the considerations that count for and against her holding an attitude and either hold this attitude or refrain from holding it depending on which response those considerations make appropriate" (Portmore 2019, 96). Given the appeal to a suitably wide range of counterfactual conditions, it seems the rational capacities view will, like the reasons-responsiveness view, fail the binary requirement.

Yet, as mentioned at the outset, the attitudes and actions of cognitively unsophisticated animals are areasonable. To avoid mistakenly classifying their actions or attitudes as assessable for reasonableness, we cannot treat the elements of their psychologies that cause these actions or attitudes as sufficient for assessability. We should thus accept the

> *Sophistication requirement*: Whether something is assessable for reasonableness depends on a factor that does not overlap with the causal chain leading to the attitudes had or actions performed by cognitively unsophisticated animals.[13]

This requirement demands that we find a feature that either is present in the causal chain when assessable for reasonableness but absent from the causal chain of cognitively unsophisticated animals, or is absent from the causal chain when assessable for reasonableness but present in the causal chain of cognitively unsophisticated animals. What motivates this requirement is the thought that if we had a perfect overlap, we'd get an extensionally implausible theory. Our theory would assess too much.[14]

The sophistication requirement rules out certain permissive views. Consider, for example, the view that holds that something is assessable for reasonableness if and only if it is intelligent—governed by the contents of one's thoughts in ways that make sense.[15] This view fails the sophistication requirement. The attitudes and actions of cognitively unsophisticated animals are often intelligent.[16] The intelligence view thus mistakenly claims that their attitudes and actions are assessable for reasonableness.

In light of the sophistication requirement, cognitively demanding attitudes might seem like the obvious place to turn. For example, it might be thought that we could satisfy this requirement by turning to attitudes that represent their contents under a normative or evaluative guise, or by turning to beliefs with normative or evaluative content.[17]

[13] It might be better to claim that cognitively unsophisticated animals rarely or only in unusual circumstances have attitudes and actions that are assessable for reasonableness. Incorporating this qualification would then give us: Whether something is assessable for reasonableness depends on a factor that does not *normally* overlap with the causal chain leading to the attitudes had or actions performed by cognitively unsophisticated animals. Adding this won't change the coming argument. I will thus stick with the unqualified formulation.

[14] For further discussion, see Regan (2003, 653–654).

[15] This view is inspired by Sidgwick (1907). He held that, "Moralists of all schools, I conceive, would agree that the moral judgments which we pass on actions relate primarily to intentional actions regarded as intentional" (1907, Bk. III, chp. 1, §2). The gloss on intelligence is borrowed from Dretske (1993).

[16] For a related discussion, see Kagan (2002, 113–114) and Tomasello (2022, chp. 4).

[17] This idea is inspired by Watson (1975, 215). If we followed Watson (1975), the relevant attitude would be some kind of evaluative judgement or belief. Alternatively, we could follow Shah (2008) or Gregory (2021) and treat the relevant attitude as some kind of normative judgement or belief. Finally, we could shift the normative or evaluative element from the content to the attitude. Yet moving it into the attitude it is strategy usually used to make the account less cognitively demanding; see, for example, Block (2023, 211). For more on the attitude/content distinction, see Tenenbaum (2021) and Kriegel (2022).

However, I will now argue that no attitude, regardless of how cognitively demanding, could alone be sufficient for being assessable for reasonableness. To see why, consider

> **One-off manipulation:** A neurologist has implanted a device into your brain. This device directly causes you to believe that you ought to lock out your roommate. This belief with normative content directly causes you to lock out your roommate.

Despite resulting from a cognitively demanding attitude, your locking out your roommate is areasonable because of the dubious origins of the causally relevant belief. And dubious origins are not limited to manipulation. They also include tumors, swift blows to the head, certain mental illnesses, and the like.

But what precisely makes the origins of attitudes and actions dubious in a way that impugns assessability? One suggestion appeals to the idea that these attitudes and actions are psychological anomalies. Given their origins, they do not fit with the rest of the agent's psychological profile. She is alienated from them in the sense that she would disavow them upon reflection.[18]

Though there is something admittedly attractive about this suggestion, it cannot be the whole story. For the problem remains even if the attitudes and actions are a perfect fit with the rest of the agent's psychology. Consider

> **Complete manipulation**: Your complete psychological profile—all your beliefs, desires, intentions, and so on—were directly caused by a device implanted into your brain by a neurologist.

It is not the case, because of the totalizing nature of the manipulation, that you would upon reflection disavow any particular attitude. There is only identification, no alienation. And yet, due to being directly caused by the device, your beliefs, desires, intentions, and the like are not assessable for reasonableness.

If alienation is not the real source of the problem, what is? The answer, I believe, is that in all of these cases the causal sequence leading to the attitudes and actions bypasses any reasoning. This answer is suggested by the fact that in each of our cases "directly" needed to be inserted to get the desired result. We needed this qualification because, without it, it would be possible for you to arrive at these attitudes and actions indirectly through reasoning. And if reason got involved, the verdict would cease to be obvious. This answer can also explain why appeals to alienation have some ring of plausibility to them. Reason unifies your psychology.[19] So in one-off cases an attitude with dubious origins will usually fail to mesh. But disavowal upon reflection is merely a symptom. The root of the problem, which is what makes turning to whether you'd endorse the attitude upon reflection attractive, lies with the fact that reason never got a chance to sign-off on it.

We can test whether this diagnosis is correct by considering a case modified from Dennett (1984, 65):

[18] This suggestion can be found in McHugh (2013, 2017).

[19] For more on this sort of unification, see Korsgaard (2009, chp. 4) and Tomasello (2014, 118).

Indirect manipulation: A well-informed, truthful oracle indirectly manipulates you by bombarding your ears with lucid and accurate warnings about your roommate, made all the more irresistible by the citation of all the evidence in their favor and a frank account of the entire evidence-gathering operation. You reason from the oracle's testimony to the conclusion that you ought to lock out your roommate. This belief directly causes you to lock out your roommate.[20]

In this case, it seems that your belief and action are assessable for reasonableness. This lends credence to our diagnosis. What makes the cut between dubious and non-dubious origins is reasoning. If your reasoning is not involved, the origins are dubious such that assessments of reasonableness are out of place. If your reasoning is involved, the origins are non-dubious, inviting assessments of reasonableness. We thus are led to the

Reasoning requirement: Something is assessable for reasonableness only if it is caused in the right way by reasoning.[21]

This requirement explains our verdicts in one-off manipulation and complete manipulation. Since reasoning is a causal process, it is consistent with causal sequence exclusivity.[22] Moreover, the account of what makes origins dubious—that reasoning has been left out of the causal story—seems well-positioned to account for what is missing from the psychologies of cognitively unsophisticated animals. Since cognitively unsophisticated animals cannot reason their way to their attitudes or actions, their attitudes and actions are not assessable for reasonableness.[23] As far as reasoning goes, they are in the same position as you in complete manipulation. The reasoning requirement offers a neat, unified diagnosis.

This requirement is crucially important for the coming argument, so it is worth pausing here to fully showcase its appeal. We have already seen that the reasoning requirement properly accounts for why cases involving dubious origins undermine assessability. But other cases—one's not involving dubious origins—pose problems for theories in normative ethics and epistemology. These cases reside at the edge of the domain of reason. Applying the reasoning requirement is illuminating.

We can start with actions. Here are three cases:

Button: An undetectable button is in front of you, well within reach. Pressing it would confer a great benefit. Yet, because it is undetectable prior to its being pressed, nobody knows of the existence of this button. Its existence could only be known after it was pressed. You reach and press the button.[24]

[20] I've retained most of Dennett's original wording.

[21] Williamson (2000, 8) suggests something similar when he writes, "If the causal explanation of the action cited only mental states immediately preceding the action, it would omit those on which the deliberation was based, and thereby miss the rationality of the action."

[22] This is a widely accepted idea. See, for example, Broome (2013).

[23] This is a bit too quick. What I should say is that cognitively unsophisticated creatures cannot engage in the kind of reasoning needed for assessability. I'll return to this point below.

[24] This sort of case often shows up in the objectivism/perspectivism debate. See, for example, Jackson (1991), Lord (2015), and Way and Whiting (2017).

Cure: You are sitting at your computer. You do not know the cure for cancer. But you know that sending an email with the cure to the World Health Organization would confer a great benefit. You know how to email the World Health Organization. You know how to type each of the words that would, if you put them together in the right order, amount to the cure for cancer. But you don't know how to put them together in the right order, because you don't know the cure for cancer. You type out the cure and send it.[25]

Unthinkable: You are a person of great integrity. You have plans, projects, and loyalties around which you have built your life. You are now placed in a situation where you know that the performance of a certain act (e.g., killing) would confer a great benefit. But this killing, given your commitments, would compromise your integrity at the deepest level. You kill.[26]

In each of these cases, it's questionable whether pushing the button, typing out the cure, or killing is assessable for reasonableness. But the reasoning requirement offers plausible guidance. It depends on whether we think it possible that you reasoned to the action in question. The question "Is pushing the button, typing out the cure, or killing assessable for reasonableness?" turns into the question, "Could your reasoning have caused these actions in the right way?"

Of course, we won't be able to give an answer until we have a more precise account of the kind of reasoning needed for assessibility and what it takes to be caused in the right way. Nevertheless, whether reasoning could have caused these actions in the right way seems to be the place to look. Notice that, as we move through the cases from Button to Cure to Unthinkable, our confidence that the actions are areasonable decreases. In Button, given that the button is undetectable, it is obvious that pushing is areasonable. In lockstep with the reasoning requirement, this is obvious because it is obvious that you couldn't have reasoned your way to pushing; you could have only push by accident. And it seems nearly as obvious that typing the cure to cancer would have bypassed your reasoning or was caused by reasoning in a particularly deviant fashion. But it is less obvious that you were unable to reason in the right way to the performance of an integrity compromising action, which is why it is less obvious that doing what was unthinkable to you is areasonable. Just as the reasoning requirement predicts, the less sure we are that you couldn't have reason in right way to an action, the less sure we are that it is areasonable.

We can next turn to attitudes. Here are three cases:

Pascal's Wager: Your available evidence overwhelmingly suggests that God does not exist. But you are offered a very favorable cost-benefit ratio for believing that God exists over withholding or disbelieving. You form the belief that God exists.

Toxin puzzle: You are offered a great benefit if you intend, at t_1, to drink a mild toxin at t_2. You know that if you drink the toxin at t_2 you'll feel sick at t_3.

[25] This sort of case is often appealed to in discussions of cognitive limitations and the response constraint. See, for example, Dorsey (2013) and Portmore (2019, §1.1.9).

[26] This case is inspired by Williams (1973).

You also know that you do not need to drink the toxin at t_2 to receive the benefit. The benefit attaches to having the intention at t_1, not executing it at t_2. You form the intention at t_1.[27]

Self-fulfilling belief: Your available evidence overwhelmingly suggests that you will not pass the upcoming exam. But if you believe that you will pass the exam, this will raise your self-confidence. And with this psychological boost, you will end up passing the exam. You form the belief that you'll pass.[28]

Are these attitudes assessable for reasonableness? The reasoning requirement tells us what we need to find out: Could these attitudes have been connected in the right way to reasoning? If it is impossible for these attitudes to have been connected to reasoning in the right way, then they lie outside the domain of reason. And, as before, the higher our confidence that you couldn't have reasoned in the right way to the attitude, the higher our confidence that the attitude is areasonable. We are quite sure that you couldn't have reasoned your way to a belief in God in Pascal's Wager, and so we are quite sure that, though this belief is good to have, it is areasonable. We are fairly confident that you could not have reasoned your way to the intention in the toxin puzzle, and so we are fairly confident that this intention is areasonable. But we are unsure whether you could have reasoned your way to the belief that you will pass the upcoming exam in self-fulfilling belief, and so we are unsure whether or not it is assessable for reasonableness.[29] Again, this correlation lends credence to the reasoning requirement.

We can make similar remarks about whether only attitudes and actions are assessable for reasonableness. Parfit (2011, 53–54), for example, claims that liking or disliking certain sensations is areasonable.[30] On the plausible assumption that it is impossible to non-deviantly reason to liking or disliking certain sensations—these states simply befall us—the reasoning requirement vindicates this claim. But Parfit also claims that admiring certain artistic works is areasonable. This claim, as Parfit admits, is controversial. The reasoning requirement tracks these varying levels of controversy. It is uncontroversial to hold that we cannot reason to likes or dislikes, so Parfit's claims about them are uncontroversial. It is controversial to hold that we cannot reason to admiring certain artistic works, so Parfit's claims about them are controversial. As before, to settle this matter, we would need an account of what cannot be caused in the right way by reasoning. But it should be clear from what has been said already that the reasoning requirement is on the right track extensionally.

Despite its appeal, it might be thought that this requirement excludes too much. Consider, for example,

Blindsight: Your right occipital lobe was surgically removed. When a stick is held up in your blinded field either horizontally or vertically, you directly form beliefs about the stick's orientation. Though you have no idea why you hold

[27] This case is from Kavka (1983).

[28] This case is from Foley (1991).

[29] The same remarks apply to epistemic trade-off cases. For discussion of such cases, see Berker (2013, 363–365) and Joyce and Weatherson (2019).

[30] For a related discussion, see Scanlon (1998, 20–21). Scanlon focuses on feelings such as being tired or hungry and other states such as being distracted.

these beliefs—their formation bypasses your reasoning—they are accurate. You are completely unaware that you can form accurate beliefs in this way.[31]

According to the reasoning requirement, your belief about the stick's orientation, because it is disconnected from your reasoning, is areasonable. Yet some may hold that, because you lack access to any evidence for this belief, it is unreasonable. Others may hold that, because your belief is accurate, it is reasonable. Although these views disagree about the verdict, they nonetheless agree that the belief is in the domain of reason.

Cases like Blindsight require careful treatment. For the resistance to the reasoning requirement is, I believe, driven by taking an atemporal perspective on the assessability of the belief, overlooking how assessability might change over time. This oversite is an understandable mistake, produced by treating assessability for attitudes as working just like assessability for actions.

In considering whether your belief in Blindsight is assessable, it is instructive to consider normal perceptual beliefs. In many ways, these work like your belief in Blindsight. When you perceive, the perception directly—without reasoning being involved—causes you to believe. Hence, according to the reasoning requirement, perceptual beliefs are areasonable at their inception.[32] Here too, some may balk at this suggestion. But remember, this verdict only holds for newly formed perceptual beliefs. And, for most of us, that is not the end of the story. When it comes to perceptual beliefs, it is common to hold that we proceed in a two-step fashion: Initially, we form beliefs automatically as a sort of psychic reflex. Later, we sustain them through reasoning.[33] Weatherson (2008, 556) suggests that we think of this process as analogous to how security works at a shopping mall: Everyone is allowed in (you believe everything you see) but security (reason) removes those it regards as undesirable.[34]

We can now explain away the resistance to the reasoning requirement. We can claim that when formed automatically—at the first-step—blindsight and perceptual beliefs are areasonable. This claim is consistent with the reasoning requirement. But we can also claim that, once reasoning gets involved—at the second-step—these

[31] For a succinct overview of this phenomenon, see Fish (2010, 63). One may worry that blindsight of the sort described here does not yield beliefs. If so, one can modify the example to superblindsight. On Block's version: "The superblindsighter spontaneously says 'Now I know that there is a horizontal line in my blind field even though I don't actually see it.' Visual information of a certain limited sort (excluding color and complicated shapes) from his blind field simply pops into his thoughts in the way that solutions to problems we've been worrying about pop into our thoughts, or in the way some people just know the time or which way is North without having any perceptual experience of it" (2002, 211).

[32] It is worth stressing that the claim made here only concerns one type of status a belief might have: reasonableness. It may be that perceptual beliefs have some other epistemic status—e.g., the sort of entitlement (warrant without reasons) defended by Burge (2020). Since perceptual beliefs are usually the result of properly functioning perceptual capacities, they are paradigmatic examples of beliefs that are epistemically warranted. I am here assuming that the objector is not confusing being areasonable with lacking all epistemic good-making features. If this confusion is what is driving the objection, then it can simply be dismissed. I thank Nikolaj Jang Lee Linding Pedersen for suggesting this point.

[33] For further defense of this two-step process, see Scanlon (2007, 90) and Raz (1999, 11).

[34] Weatherston is following the work of Gilbert et al. (1993).

beliefs can be assessable for reasonableness. This claim is consistent with the reaction that blindsight and perceptual beliefs are assessable.

That assessability changes over time can be supported by contrasting blindsight with

> **Iterated blindsight:** This case is just like Blindsight except researchers have been testing you over and over for weeks. After each test, they show you the stick's orientation. As a result of repeated confirmation, you form the belief that your blindsight beliefs are usually accurate. On your 1000th test, the researchers tell you they will not show you the stick afterward. During this test the stick is held up in your blinded field vertically, and you directly—bypassing any reasoning—form the belief that the stick is held vertically. You then reason that since you have a very good past track record and believe that the stick is held vertically, the stick is likely held vertically. This reasoning sustains your belief that the stick is held vertically.

In this case, the two-step process is on full display. And it should be clear that there is an important difference in assessability between the belief when initially formed and the belief when sustained by reasoning.

Those that have the reaction that the reasoning requirement is too restrictive fail to keep these steps distinct. But keeping these steps distinct is crucial. Blindsight beliefs and perceptual beliefs are, in terms of the causal sequence leading to their initial formation, not dissimilar to the causal sequence leading to the formation of beliefs for cognitively unsophisticated animals. So if we argue that newly formed blindsight beliefs and perceptual beliefs are assessable for reasonableness, then we would, by parity of argument, be forced to hold that the beliefs of cognitively unsophisticated animals are also assessable. If, however, we hold that initially these beliefs are areasonable and only after they have been sustained through reasoning are they assessable, we can avoid this result.[35]

We can now return to the mistake that seems to be driving this objection. Notice, as the last example makes clear, that attitudes and actions are disanalogous. As persisting states, attitudes can be initially formed without reasoning but then later sustained by reasoning. Actions cannot be. The two-step process that can occur with blindsight and perceptual beliefs cannot occur with actions. Without the opportunity to be sustained through reasoning, the reasoning requirement predicts that actions will differ from blindsight and perceptual beliefs: If an action is areasonable, it is always areasonable. For example, consider

> **Recoil**: Late at night, you are reading next to a window. Suddenly, a face appears, staring at you through the glass. You recoil.[36]

Is your recoiling assessable for reasonableness? The answer seems to be No. Your recoiling was an instinctive reaction—a reflex hardwired by nature. But, unlike

[35] Scanlon suggests something similar when he writes, "Not only perceptual beliefs, but many other attitudes as well arise in us unbidden, without conscious choice or decision. Nonetheless, as continuing states these attitudes are 'up to us'" (1998, 22).

[36] This case is modified from Anscombe (1963, §5).

beliefs, you cannot sustain a recoil through reasoning. Actions are performed, not sustained.[37] Once performed, they are over and done with. You can go on to do something else that is very much like a recoil through reasoning. But that is not sustaining the recoil; it is performing a different action—holding yourself in a recoiled position, say. Once completed, there is no way for your reasoning to reach back to the initial event. Going forward, you are never sustaining; you are always doing something new. Accordingly, the reasoning requirement deems your recoiling areasonable without any chance of later entering the domain of reason. That seems to be the correct result.

The difference between attitudes and actions we've been considering is easy to overlook. This supports the suggestion that the resistance to the reasoning requirement is animated by mistakenly taking an atemporal perspective. For it is tempting to think that what holds for assessability of actions must also hold for the assessability of attitudes. But, by noticing that attitudes can be sustained through reasoning while actions cannot, we can resist this temptation.

Now for a more pressing objection. Reasoning admits of varying degrees of sophistication. On one end of the continuum, we have the inferences involved in subpersonal information processing. At the other end, we have what Boghossian (2018, 56–57) calls fully explicit reasoning. In between, we have the sorts of inferences performed by animals navigating their environment and the logic conforming inferences performed by children.[38] Thus it appears we face a smooth continuum with no non-arbitrary cut off. So it might seem that, unless we accept that every level of sophistication counts, we will fail to satisfy the binary requirement. But an account that accepts every form of reasoning (or inference) would violate the sophistication requirement. Thus, in appealing to reasoning, we seem to confront a dilemma: Either we fail the binary requirement or we fail the sophistication requirement.

We can escape this dilemma. The dilemma is predicated on the assumption that there is only one non-arbitrary location on the continuum from sub-personal inferring to fully explicit reasoning. But this assumption is mistaken.

Reasoning that involves one having thoughts about relations of support—thoughts about evidence, reasons, warrant, justification, implication, consequence, and the like—are importantly different from mental processes that do not involve such thoughts. Many hold, for example, that genuine reasoning must involve representations of support relations:

> Boghossian (2014, 4): *S*'s inferring from *p* to *q* is for *S* to judge *q* because *S* takes the (presumed) truth of *p* to provide support for *q*.
> Broome (2019, 32): Reasoning is a mental process through which you acquire a new attitude—the "conclusion attitude"—on the basis of attitudes you

[37] For related a discussion of the difference between attitudes and actions, see Setiya (2008) and Hieronymi (2009).

[38] As Darwin notes, "Few persons any longer dispute that animals possess some power of reasoning. Animals may constantly be seen to pause, deliberate, and resolve. It is a significant fact, that the more the habits of any particular animal are studied by a naturalist, the more he attributes to reason and the less to unlearnt instincts" (2009 [1871], 46).

already have—the "premise attitudes." It is very natural to think that, if a process is to be genuinely reasoning, you must believe that the conclusion attitude is linked to the premise attitude in some way that makes it appropriate to have the conclusion attitude on the basis of the premise attitudes. I adopted this natural view in my book *Rationality Through Reasoning*; I assumed you must have a "linking belief," as I call it.[39]

Pettit (2016, 3374): [R]easoning does not lead you to form or confirm this conclusion-belief by brutely jogging you into that state; it operates only as a causal result of a "linking-belief": a belief that the premises (say, "*p*," "*q*," and "*r*") imply the conclusion (say, "*t*") […] In the exemplar case of express or explicit reasoning, a word like "so" will mark the appearance of this conclusion-belief in the manner characteristic of reasoning as distinct from any other belief-generating process.

Neta (2013, 403): *S* infers *q* from *p* = *S* judges: *p* and therefore *q* (where the contextually salient explanatory relation is the relation of doxastic justification, or its practical analog).

Shah (2006, 486): Deliberation, or reasoning, is the process in which agents recognize reasons, and then φ on the basis of this recognition.

Burge (1996, 98–99): Critical reasoning is reasoning that involves an ability to recognize and effectively employ reasonable criticism or support for reasons and reasoning. It is reasoning guided by an appreciation, use, and assessment of reasons and reasoning as such. As a critical reasoner, one not only reasons. One recognizes reasons as reasons. […] A non-critical reasoner reasons blind, without appreciating reasons as reasons. Animals and small children reason in this way.

We do not need to accept that all genuine reasoning must involve metacognitive representations of support relations.[40] All that's needed is the far less controversial claim that there is a non-arbitrary difference between reasoning that involves representing support relations and reasoning that doesn't. Let us, following Burge, call the former *critical reasoning* and the latter *non-critical reasoning*.

Deploying this distinction, we can claim that, even if reasoning admits of degrees of sophistication, something is assessable for reasonableness only if it is caused in the right way by critical reasoning.[41] This claim avoids running afoul of the binary requirement. Insofar as our explanation invokes the fact that this reasoning involves recognizing relations of support, we do not face the worry that our answer to the

[39] Broome goes on to reject his earlier view. Notice that, even if he is correct in rejecting this view, he is rejecting a claim that is much stronger than the claim I need for the argument to go through. We only need the weaker claim that there is a non-arbitrary break between reasoning that does, and reasoning that does not, involve representing support relations. Nothing Broome says casts doubt on there being such a break.

[40] For a powerful argument that propositional reasoning does not require metacognition of this sort, see Burge (2010b, 54–56).

[41] This helps explain the reasons-responsiveness view's appeal. If critical reasoning is needed for assessability, then we have as part of the causal chain the recognition and assessment, through a mechanism, of reasons. As McHugh notes in explaining his version of the reasons-responsiveness view: "[W]e exercise attitudinal control when we revise our attitudes by responding directly to reasons. How do we do this? In the paradigm case, we do it by reasoning" (2017, 2756).

explanatory question will apply with equal force to less sophisticated forms of reasoning.[42] That is, an explanation which appeals to the presence of thoughts about support relations in critical reasoning will not carry over to non-critical reasoning.[43] We also avoid running afoul of the sophistication requirement. Cognitively unsophisticated animals are non-critical reasoners.[44] Claiming that critical reasoning marks the beginning of the domain of reason thus allows us to successfully navigate the dilemma, satisfying both the binary requirement and the sophistication requirement.

We are now in a position to put forward an answer to the possession question, namely, the

> *Critical reasoning view*: Something is assessable for reasonableness if and only if it is caused in the right way by critical reasoning.[45]

This view satisfies all our requirements. That speaks in favor of the view. But a final extensional adequacy worry remains.

Recall that we were looking for the part of the causal chain that converts the areasonable into things that possess the property of being reasonable or unreasonable. The critical reasoning view claims that this conversion is accomplished through reasoning that involves representations of support. Here is where the worry arises. For it might be thought that the nature of action and the nature of critical reasoning are such that the critical reasoning view renders all actions areasonable.[46]

[42] This helps explain the judgement-sensitivity view's appeal. If critical reasoning is needed for assessability, then the judgement-sensitivity view is getting part of the story right. Assessability for reasonableness always involves something in the vicinity of a judgment that there is sufficient reason to hold the attitude or perform the action.

[43] A few implications are here worth noting. If a child and an adult both believe that *p* but the child's belief was arrived at by non-critical reasoning while the adult's was arrived at by critical reasoning, only the adult's belief is assessable for reasoning. If an adult has the capacity to engage in critical reasoning and so could have arrived at a belief that *p* through such reasoning, but instead ends up believing p because of wishful thinking or some other unreliable process, the belief would be areasonable. Relatedly, it may be that on one occasion an adult reaches a belief that *p* through non-critical reasoning and so the belief is not assessable for reasonableness. On another occasion the same adult comes to the same belief—the belief that *p*—but on the basis of inference carried out through critical reasoning and so is assessable. Some might find these implications unintuitive. Yet, I not only find them intuitively plausible but also appealing, since they seem to flow naturally from the arguments in favor of Causal Sequence Exclusivity and the Sophistication Requirement (irrespective of the details of our account). I thank Nikolaj Jang Lee Linding Pedersen for pressing me to make this explicit and for some of the wording used here.

[44] Or, in more qualified terms, they are not normally critical reasoners.

[45] Notice that, even if we assume Swampman (Davidson 1987) has attitudes and performs actions, all of these attitudes and actions will at inception be areasonable, because there has not been time for critical reasoning to cause them. The same goes for Instant Agents (Timmerman and Cohen 2016). Additionally, reliably formed true beliefs that lack a connection to critical reasoning will also be areasonable. So, for example, Clairvoyant Norman's (BonJour 1985, 41) beliefs about the location of the President or Mr Truetemp's (Lehrer 1990, 163–164) beliefs about the temperature fall outside the domain of reason. These beliefs were not formed or sustained by critical reasoning. I take these implications to be plausible.

[46] For overviews of the ideas driving this objection, see Smith (1994), Dancy (1993), and Sinhababu (2017).

This objection is, I believe, the most serious threat to the critical reasoning view. It is thus worth considering in some detail.

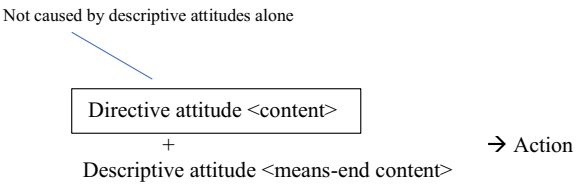
The objection begins with a familiar way of dividing up the attitudes required for an event to be an action. We have *descriptive attitudes* which, similar to the indicative mood in speech, host their content as being true. They have truth-conditions. And we have *directive attitudes* which, similar to the imperative mood in speech, host their content as being goals or ends. They have fulfillment-conditions.[47] Given their different roles, we need both to cause an action. Directive attitudes tell us were to go; descriptive attitudes supply a map with a path to get there. Both are required because, without directive attitudes, we wouldn't have an end; and, without descriptive attitudes, we wouldn't have the means to achieve it. Weakening this gives us

> *Premise 1*: For an event to be an action, its etiology must include directive attitudes.[48]

Not only does the difference between the descriptive and directive support premise 1, it also supports

> *Premise 2*: The directive attitudes that must be included in an event's etiology for it to be an action cannot be caused by descriptive attitudes alone.

Taken together premise 1 and premise 2 tell us that, for an event to be an action, its ultimate attitudinal etiology cannot be exclusively descriptive. The thought behind this claim is that descriptive attitudes represent the ways things are disinterestedly, neutrally. By themselves, they cannot cast the action in a favorable light. They are motivationally inert. By contrast, directive attitudes are motivationally efficacious and capable of rationalizing. This makes them indispensable for action. Moreover, this difference in rationalizing and motivational power puts these directive attitudes ultimately outside of the reach of descriptive attitudes. Of course, one might come to a new directive attitude (e.g., the desire that p) partly on the basis of a descriptive attitude (e.g., the belief that p is a means to *q*). But to do so, the descriptive attitude must have been combined with an already held directive attitude (e.g., the desire that *q*). Hence, when we trace back the causal chain of any action, we will always find at least one directive attitude that is not caused by descriptive attitudes alone. Premise 1 and premise 2 combine to give us the following picture:



[47] Here I am following Millikan (2004, chp. 6). Her terminology is to be preferred to cognitive/non-cognitive or belief-like/desire-like. It is preferable because, unlike those alternatives, it allows for talk of directives to be shifted from attitude to content. This will matter below.

[48] This premise is usually taken to be part of the Standard Story of Action, see Davidson (1980). For discussion and defense, see Smith (2012, 2013).

This picture is then combined with the idea that the premise-attitudes of critical reasoning are descriptive attitudes.[49] This gives us

Premise 3: The premise attitudes and the conclusion attitudes of critical reasoning are exclusively descriptive.

The most interesting argument for premise 3 focuses on the nature of critical reasoning. Unlike other forms of reasoning, critical reasoning comes with an extra commitment. It must involve representations of support. And this has an important implication if we accept, as seems plausible, that support can only be provided by things with truth-conditions. If representations of support must involve contents hosted as true, then the premise-attitudes in critical reasoning must be exclusively descriptive.[50] And, if the premise-attitudes are all descriptive, and descriptive attitudes alone cannot cause directive attitudes, then the conclusion-attitude of any bit of critical reasoning must also be descriptive. Putting these three premises of the argument together takes us to the

Conclusion: Critical reasoning, by itself, cannot cause actions.[51]

From this conclusion we seem licensed to infer, if the critical reasoning view is true, no actions are caused in the right way to be assessable for reasonableness.[52]

Since I believe there might be something to the argument that critical reasoning trades in descriptive attitudes alone, I will grant premise 3.[53] Still, we can resist the conclusion. Although not dispositive, there is a good case to be made that either premise 1 or premise 2 is false.

To make this case, we need to get clearer on the thought behind premise 1 and premise 2. That idea, recall, is that descriptive attitudes and directive attitudes are very different. So different that in action we must have a least one directive attitude, and so different that descriptive attitudes alone cannot cause the directive attitudes required for action. What is this deep difference? It is, on the standard view, a difference in direction of fit.[54] Platts succinctly captures the driving idea:

> Beliefs aim at the true, and their being true is their fitting the world; falsity is a decisive failing in a belief, and false beliefs should be discarded; beliefs should be changed to fit with the world, not vice versa. Desires aim at realisation, and their realisation is the world fitting with them; the fact that the indicative content of a desire is not realised in the world is not yet a failing in the desire, and not yet any

[49] This claim follows from a cognitivist account of reasoning, which, if true, would cover critical reasoning as a species. For the classic defense of this account, see Hume (1960, T. 2.3.3).

[50] Something in the spirit of this argument is discussed in Staffel (2019, 59–60).

[51] This, of course, is Hume conclusion (1960, T 3.1.1.8) restricted to critical reasoning.

[52] In point of fact, however, there is room to hold that so long as the descriptive attitudes are caused by critical reasoning the action is assessable for reasonableness. What's more, even if we accept that the premise-attitudes of critical reasoning are exclusively descriptive, this does not, for reasons which will be made clear below, imply that conclusion-attitudes cannot be directive.

[53] For a more liberal account of what we can reason to, see Drucker (2022). And for an account that allows for different attitudes to enter in as premises by way of marked contents, see Broome (2013, chp. 14). I tentatively hold that, when it comes to critical reasoning, these accounts cannot be made to work.

[54] For discussion and critique of the standard view of direction of fit, see Archer (2015) and Frost (2014). The details of the metaphor won't matter for the coming response. So I am happy to go along with the standard understanding.

> reason to discard the desire; the world, crudely, should be changed to fit with our desires, not vice versa. (1997, 256–257)

As a descriptive attitude, a belief that the cat is on the mat aims to fit the world.[55] If the cat is not on the mat, your mind needs to be revised to align with the world. A desire to pet the cat, as a directive attitude, aims to have the world fit it. If you are not petting the cat, the revision needs to be made to the world not your mind. This difference in direction of fit explains two important features of directive attitudes: their ability to rationalize and motivate.

Action demands rationalization and motivation. The agent needs to have an end to cast the action in a favorable light such that the doing of it makes sense from the agent's point of view. And relatedly, the agent needs to be poised to bend the arc of history toward this end—to be disposed to make interventions on the world to achieve the goal. A belief that the cat is on the mat does not cast petting in a favorable light and it is impotent to move the agent. Directive attitudes (with world-to-mind direction of fit)—e.g., the desire to pet the cat—can rationalize and motivate. And descriptive attitudes (with mind-to-world direction of fit) cannot by themselves get a grip on directive attitudes. The belief that the cat is on the mat cannot, without tapping into some antecedently held desire, produce a new desire. It has the wrong direction of fit to do so.

This explanation of the deep divide between descriptive attitudes and directive attitudes is appealing.[56] But its appeal fades once we notice descriptive attitudes can have directive content.[57] For any attitude normally classified as directive, we can find a close descriptive counterpart. Consider the following pairs: Intending that p and believing that *p* is choiceworthy (or is an end worth pursuing). Valuing that *p* and believing that *p* is valuable. Desiring that *p* and believing that *p* is desirable (or is good, or ought to be the case). Fearing that *p* and believing that *p* is fearsome (or dangerous). Admiring that *p* and believing that *p* is admirable. Respecting that *p* and

[55] For discussion of the direction of fit metaphor as it relates to the argument under discussion, see Smith (1994, chp. 4).

[56] If we had started with our evolutionary history rather than a metaphor, we would be even less inclined to find this divide appealing. According to a plausible story advanced by Millikan (1995, 2004, 2017), all representation began with "pushmi-pullyus," which are representations that are simultaneously descriptive and directive. They have both truth conditions and fulfillment conditions. For example "the reflex that withdraws the hand from something unexpectedly hot, are mediated by P-Ps […telling] what part of the body is exposed to something too hot and directs withdrawal of that part" (2004, 157). Humans have detached these two parts; we can have directive attitudes and descriptive attitudes. But how we got there suggests that this detachment is not clean. As Millikan explains, "Beginning with minimally articulate P-P representations, the evolution of inner representations seems likely to have paralleled evolution writ large. First, representations have become more articulate, so that more and more of what they represent is represented explicitly. More complex functions are then built up of out of more specialized functions of the articulated parts. Then ways to perfect these more specialized functions somewhat independently have developed, sometimes by the development of new generate and test procedures. These articulated specialized functions are then recombined and reintegrated in new ways. The general strategy involved—disassemble, tune the parts separately and recombine—is typical of evolutionary developments more generally" (2004, 172). If Millikan's story is on the right track, then, given their common ancestor in pushmi-pullyus representations, it would, contrary to Humean orthodoxy, be in fact surprising if descriptive attitudes and directive attitudes were distinct existences. For a different but related discussion claiming that the most primitive states that motivate action are both descriptive and directive, see Shea (2014).

[57] For discussion of this point, see Gregory (2021, §1.3–1.4).

believing that *p* is worthy of respect. Regretting that *p* and believing that *p* is regrettable. And so on.[58]

With these pairs in view, the divide between the descriptive and the directive looks less deep. Descriptive attitudes with directive content makes the prospects for action explanation in ultimately descriptive terms look far more promising. If we can show that the ultimate attitudinal etiology of actions can be exclusively descriptive, then we must reject premise 1 or premise 2. And we can show this by showing any of the following: Descriptive attitudes alone cause actions; descriptive attitudes alone cause composite attitudes—attitudes that are both directive and descriptive—which cause actions; or descriptive attitudes alone cause pure directive attitudes which cause actions. These disjuncts are compatible. So below I'll make the case that each is a viable possibility. But it is worth bearing in mind that it only takes the truth of one disjunct to rebuff the objection.

Start with the first possibility, returning to the case of indirect manipulation. Recall, in this case, you reasoned to a belief with normative content which directly caused you to lock out your roommate. So we were assuming that descriptive attitudes alone, some with directive content, can cause actions. That is, we were assuming the following picture:

Descriptive attitude <directive content>
\+ → Action
Descriptive attitude <means-end content>

I take this case to be plausible as described. But the defender of the objection will press that this case is underdescribed. Background directive attitudes were in fact there but, as a conversational shorthand, quietly omitted. After all, you were given warnings about your roommate. Thus it is tempting to suppose that other directive states—e.g., fear of being hurt, the desire not to be hurt, and the like—were present. We need a better case.

Consider Kant's Friend of Humanity. At first, he possesses directive attitudes aimed at helping others. Later, he is overcome by grief, sapping him of these attitudes. Still even when "no inclination any longer stimulates him to it, he tears himself out of this deadly insensibility and does the action without any inclination, solely from duty" (2002, Ak 4:398). This Friend of Humanity's beneficent actions are, as described, driven by descriptive attitudes alone. The belief that helping others is his moral duty—a descriptive attitude with normative content—together with other descriptive attitudes about the means to accomplish this are sufficient to move him to action. And this is not an anomalous case. Schueler offers a more mundane example: "I would say, for instance, that I had no desire to attend a meeting at my son's school the other evening. I would much rather have stayed home and read. But I did attend the meeting because I believed I had a responsibility to do so, a responsibility mostly to my son but partly to my community" (1995, 29). We can make similar remarks about cases involving descriptive attitudes with evaluative content. And, like descriptive beliefs with normative content, it seems plausible to hold that, say, a belief that something is good or worthwhile along with other means-end beliefs can move one to action.

[58] For further criticism of the appeal to direction of fit along similar lines, see Price (1989).

From the discussion of this first possibility, we can see that descriptive attitudes with normative or evaluative content can play at least one role usually assigned to directive attitudes. Recall the thought behind premise 1 was that we needed a directive attitude—we needed something with world-to-mind direction of fit—to rationalize the action.[59] Without directive attitudes, we wouldn't be able to cast the action in a favorable light such that we could rationalize it from the agent's own point of view. But, as the cases above suggest, a belief that some action is morally required or that something is a final end can play this rationalizing role.[60] And it is this ability to play this role that explains the appeal of treating typical directive attitudes—e.g., desire, intention, and valuing—as a sub-class of belief.[61]

However, one might argue that we should not lump motivation and rationalization together. Perhaps, descriptive attitudes with directive content can rationalize the action. But motivation is a separate issue. One can believe that an action is morally required, ought to be done, help achieve one's final ends and yet be left cold. What this suggests is that descriptive attitudes alone, even those with directive content, are motivationally inert. Only directive attitudes motivate. And so, without directive attitudes, the agent wouldn't move.

We should not accept this argument. If both attitudes and content can be directive, why must motivation attach exclusively to attitudes? The fact that descriptive attitudes with directive content sometimes fail to motivate shows very little. At most, it forces a minor qualification: Normally, when one has such a descriptive attitude with directive content it motivates. As Dancy writes, "the overall moral judgement 'This is what I ought to do' […] though in certain unusual circumstances it may be deprived of its normal motivational force, still has a normal motivational force to be deprived of. This idea of a normal (default) force in the overall judgement […] is surely what is gestured towards by the original intuition that there is something odd about saying 'This is wrong but I don't see that as relevant to my choice of action'" (1993, 26).[62]

Having voiced these reservations, let's grant for now that directive attitudes have a monopoly on motivation. That seems to secure premise 1.[63] But threats to premise 2 remain.

[59] I say usually, but it is worth remembering that Davidson's own list of the pro-attitudes needed to rationalize action includes descriptive attitudes with normative content. Here's the passage, "A reason rationalizes an action only if it leads us to see something the agent saw, or thought he saw, in his action—some feature, consequence, or aspect of the action the agent wanted, desired, prized, held dear, *thought dutiful, beneficial, obligatory, or agreeable*" (1980, 13; emphasis added). Williams, although this is often overlooked, also endorses the idea that descriptive attitudes with normative content are motivationally efficacious: "Does believing that a particular consideration is a reason to act in a particular way provide, or indeed constitute, a motivation to act? […] Let us grant that it does—this claim indeed seems plausible, so long at least as the connexion between such beliefs and the disposition to act is not tightened to that unnecessary degree which excludes akrasia" (1981, 107).

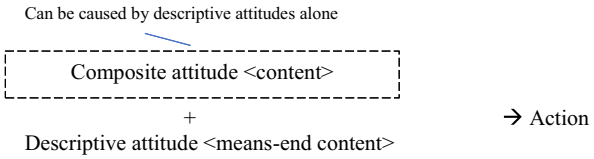
[60] Indeed some argue that only descriptive attitudes with directive content can play this rationalizing role. See, for example, Buss (1999) and Gregory (2019).

[61] For this account of desire, see Gregory (2021). For intention, see Shah (2008) or Marušić and Schwenkler (2018). For valuing, see Smith (1992). And for a cognitivist account of emotions, see Nussbaum (2001).

[62] For a related discussion, see Dreier (1990, 9–14).

[63] It would secure it if we accept that action requires motivation. If motivation is understood as some kind of feeling, then we shouldn't accept this requirement. But if we accept that motivation just means poised or disposed to act, then the requirement is far more plausible. In what follows, I'll assume the latter reading. For a discussion that takes motivation to be a psychological disposition (not feeling), see Sinhababu (2017, §2.1).

One such threat holds that some attitudes are mixed—both directive and descriptive.[64] And one can be caused to have these composite attitudes by descriptive attitudes alone. This possibility can be pictured as follows:



Cognitivist theories of emotion, which hold that descriptive attitudes with evaluative content partly constitute emotions, offer an example of this sort of view.[65] Consider pride. Pride involves descriptive elements. It involves beliefs that something is splendid or worthy of pride. Or perhaps it involves beliefs that the thing has the properties that make it the case that it is splendid or worthy of pride. And yet pride can still motivate action. This is because pride, despite being descriptively anchored, has affective and motivational elements as well. Here is Foot defending this sort of account:

> A feeling of pride is not identified like a tickle, but requires a special kind of thought about the thing of which one feels proud. Now I should say […] that it is just as bad to try to identify a feeling as a feeling of approval, whether moral approval or any other, without its particular objects as it is to try to identify pride without talking about the only kinds of things about which one can logically feel proud. (I do not mean, of course, that one would be illogical in feeling pride towards something which one did not believe to be in some way splendid and in some way one's own, but that the concept of pride does not allow us to talk like that.) Similarly for the concept of approval, though the reader will kindly excuse me from giving an account of what exactly a man must believe of those things of which he can logically approve. Anyone who doubts this point about approval should ask what it would be to have this feeling when contemplating an object one did not see as useful, beautiful, efficient or anything like that. Does it make sense to suppose that one might wake up one morning feeling approval of something believed to be an ordinary, unnecessary, unbeautiful speck of dust? (1978, 76)

According to Foot, having certain evaluative beliefs—e.g., that the thing in question is splendid—is a necessary condition on feeling proud. And it is this descriptive element which makes the mental state one of pride. Other emotions are type-identified by having a descriptive element with different directive content.[66]

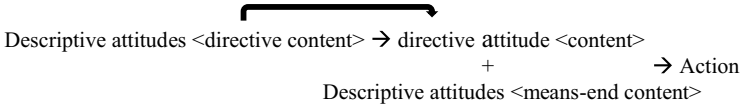
[64] For a defense of single attitudes with opposing directions of fit, see Little (1997). See also Platts (1997, 257) who holds that "all desires involve elements of belief."

[65] For an early articulation of this account, see Broad (1954). For a more recent discussion focusing on pride and guilt, see Portmore (2019, chp. 2).

[66] Ross offers a similar view of admiration. He writes, "Admiration is not a mere emotion; it is an emotion accompanied by the thought that that which is admired is good" (1939, 278).

If Foot's account is on the right track, descriptive attitudes alone have a clear path to adjusting action. Suppose you have just successfully climbed Everest. You are, motivated by pride, about to send a postcard to your friend. But as you stand in front of the post box, you reflect. You come to see the worthlessness of the activity. It is, you believe, a stupid, arbitrary goal. Modern equipment robs the challenge of expressing human excellence. And it is something that thousands of other people have already accomplished. That you believe there's nothing splendid in the climb causes your pride to vanish. Once gone, you are no longer motivated to send the postcard. Instead, you are, on the basis of these considerations, filled with regret at having wasted your efforts. As a means of expressing your regret, you crumple the card and throw it in a nearby rubbish bin.[67] As this case shows, when the cogs of cognition turn, these composite attitudes turn with them. Desire, intention, or valuing could be thought of along similar lines.[68]

The other possibility that threatens premise 2 has it that directive attitudes and descriptive attitudes are distinct, but the directive attitudes required for action can be caused by the corresponding descriptive attitudes with directive content.[69] We can picture this view in the following way:



Here is Parfit offering one way of filling in the details:

> When we come to have some belief—such as the belief that some aim is worth achieving—that might cause us to have some wholly new desire. Such a belief could not all by itself cause us to have this desire, since we would have to be such that, if we came to have this belief, that would cause us to have this desire. But this disposition may not itself be a desire. […] By giving us such beliefs, reason might motivate us without the help of any independent desire. (1997, 105)

The idea, expressed metaphorically, is that once a certain directive content is deposited into your descriptive attitude box, you are such that normally the

[67] Broome offers a similar objection to Hume: "You intend to do something, but then by theoretical reasoning you come to believe you cannot do it. You will give up your intention, since you cannot intend to do something you believe you cannot do" (2013, 292).

[68] Hybrid expressivists take moral judgment to be partly belief-like and partly desire-like. For discussion, see the essays in Fletcher and Ridge (2014). For valuing, see Scheffler (2011). For a composite account of intention, see Grice (1971). Offering a similar account, Millikan argues that intentions are pullyu-pushmi representations: "Suppose that my brain already harbors, for purposes of guiding my action, a representation of what I am definitely going to do. And suppose there is need to take this settled future into account when making further decisions about what else I can compatibly do. It would surely waste space and introduce unnecessary mechanisms for evolution to duplicate the representation I already have. Better just to use it over again as a descriptive representation as well. Notice […] that this kind of [P-Ps …] maps variations in goals directly onto the represented future world. […T]he contents of the directive and descriptive aspects of the representation are not different but coincide" (1995, 192–193).

[69] This possibility was famously defended by Nagel (1970).

corresponding content is also deposited in your directive attitude box.[70] For example, once $<p$ is desirable$>$ gets into your belief box, this causes $<p>$ to go into your desire box.[71] When you believe$<$ I ought to see to it that q forthwith$>$, this causes $<q>$ to be added to your intention box. The same holds for other descriptive-directive pairs.[72]

Here it might be objected that this conversion is mysterious. For any purported cases of descriptive attitudes causing the corresponding directive attitudes, the better explanation is that directive attitudes are operating in the background. You must, according to this objection, have a background desire—e.g., the desire to do what is worth achieving (de dicto). And it is this background desire that explains the transition from the belief that p is worth achieving to the desire that p.

Certain cases, however, escape this objection. Consider

> **Moorean goodness**: Given your moral education, the notion of an intrinsically good state of affairs is unintelligible. The only coherent evaluative notions for you are fundamentally linked to attributive goodness. However, one day, after careful critical reasoning, you conclude that there are good states of affairs. This conclusion causes you to desire that these good states of affairs obtain.[73]

In this case, a change in descriptive attitudes changes what is conceptually available and thereby changes your motivational profile. Since the relevant notion—good-simpliciter—was unintelligible for you, no descriptive attitudes concerning good-simpliciter could have been operating covertly in the background. If this case is possible, which to all appearances it seems to be, then, even if we assume that only directive attitudes motivate, we have a counterexample to premise 2.

At this point, one might try to secure premise 2 by appealing to a more radical view about motivation. On this view, certain pure directive attitudes are the sole source of all motivation, and these attitudes are causally independent from descriptive attitudes. Descriptive attitudes can only cause composite attitudes or directive

[70] The metaphor comes from Schiffer (1981, 212). For a similar idea—that we have a mechanism for depositing what is in one propositional attitude box into another—see Nichols and Stich (2003, §4.3).

[71] Here is Searle endorsing a normative version of this idea: "[P]eople can *want* to fulfill their obligations and keep their promises. Yes, but that is not like wanting chocolate ice cream. I want chocolate and I want to keep my promise. What's the difference? In the case of the promise the desire is derived from the recognition of the desire-independent reason, that is, the obligation. The reason is prior to the desire and the ground of the desire. In the case of chocolate the desire is the reason" (2001, 170).

[72] One plausible variant of this view is worth highlighting. On this view, descriptive attitude with normative content target a certain attitude. This target attitude then gets formed causally or constitutively as a result. For example, just as you might believe$<$ you ought to believe $p>$ and this causes you to believe p, you might believe$<$ that you ought to desire that $p>$ and this causes you to desire that p.

[73] This case also serves as a counterexample to Aristotle's remark that, "We deliberate not about ends but about what contributes to ends. For a doctor does not deliberate whether he shall heal, nor an orator whether he shall convince, nor a statesman whether he shall produce law and order, nor does any one else deliberate about his end" (1984, 1112b12-1113a2). This remark is in line with the objection to the Critical Reasoning View we are considering. Yet, as the above case shows, the idea that no one deliberates about ends is hard to accept. And Aristotle's own examples serve as more modest counterexamples to the view he forwards. People reason about whether to be doctors. Orators might ask themselves why they are still doing this, come to the conclusion that it is meaningless, and change professions. Our final ends do not simply befall us; we reason to them. For a discussion that makes similar remarks, see Rescher (1988, 92-94).

attitudes by tapping into these original sources of motivation, and these original sources of motivation are completely causally walled off from directive attitudes. On this picture, motivation is like gold.[74] It can be moved around, but no amount of descriptive attitude alchemy can generate it.[75]

This view has several problems. Chief among these problems is that original sources of motivation cannot be completely causally walled off from directive attitudes. Return to complete manipulation. In that case, new sources of motivation were caused by a device implanted into your brain by a neurologist. And less extreme examples are ready to hand. Hypnosis, tumors, drugs, and swift blows to the head can cause new motivationally efficacious directive attitudes.[76] And if hypnosis, tumors, drugs, and swift blows to the head can cause new motivationally efficacious directive attitudes, descriptive attitudes can exploit similar causal paths to arrive at new sources of motivation as well. In other words, the recipe for generating counterexamples to this view is straightforward: Start with descriptive attitudes alone, and then have these attitudes kick off a causal chain—no matter how wild—that eventuates in a new source of motivation.

Yet a more plausible version of this radical view is available. As before, this view holds that only certain pure directive attitudes can serve as original sources of motivation. And, while admitting that descriptive attitudes alone can cause these attitudes, it goes on to claim that descriptive attitudes can only do this in the same manner as hypnosis, tumors, drugs, and swift blows to the head. New sources of motivation are only accessible to descriptive attitudes through wayward causation. That is, no amount of descriptive attitude alchemy can alone non-deviantly generate motivation.[77] For example, the belief that *p* is worthwhile can only cause the desire that *p* in two ways: by tapping in to some other antecedently held motivationally efficacious directive attitude, or by some deviant causal route.

Though more plausible, this view still rules out the possibility that directive content can motivate. It claims Kant's Friend of Humanity, as described, is impossible. Since this case seems possible, that cast doubt on the view. It also rules out the possibility that descriptive attitudes can by themselves non-deviantly cause motivationally efficacious composite attitudes. This strains credulity. Your loss of pride and new regret after climbing Everest does not seem impossible. Nor does it exhibit any of the fortuitousness, accidentality, or flukiness characteristic of deviant causation. The same holds for the case of Moorean goodness. Admittedly, these remarks are not decisive. But it shows that adopting this view of motivation puts one in the unenviable position of showing that all of the cases discussed above are suppressing

[74] Remember we are here thinking of motivation as a disposition, not a psychologically spreadable feeling.

[75] For discussion of this line of argument, see Wallace (1990) and Williams (1981).

[76] Broome (2013, 293–294) makes similar observations.

[77] Mele (2003) endorses something close to this view. He calls it the Antecedent Motivation Theory. On this theory, "[I]n actual human beings, all motivation nonaccidentally produced by practical reasoning issuing in a belief favoring a course of action derives at least partly from motivation-encompassing attitudes already present in the agent before he acquires the belief" (2003, 89).

crucial elements, impossible, or involve deviant causation. The burden of proof lies with defender of this radical view. I see no way of meeting this burden.

We can now take stock. We have examined three possible ways in which descriptive attitudes alone could cause, directly or indirectly, an action. If even one of these possibilities is actual, the ultimate attitudinal etiology of actions can be exclusively descriptive. In fact, I think all three possibilities occur for different attitude and different content combinations. But I also think that there is room for doubt. At this point, we must compare the plausibility of at least one of these three possibilities holding in at least some cases with the plausibility of premise 1 and premise 2 taken in conjunction. The former is more plausible. We should thus cautiously hold that the critical reasoning view does not yield the implausible implication that all actions are areasonable.

We can now finally turn to what is perhaps most compelling about the critical reasoning view: how it sets us up to offer a satisfying answer to the explanatory question.

3 The explanatory question

In answering the explanatory question, we can start with a part of the critical reasoning view that demands further elaboration: being caused in the right way. This is a useful place to begin because, as we shall see, sorting out what it takes for something to be caused in the right way by critical reasoning strongly suggests a particular answer to the explanatory question.

Why does the critical reasoning view need a "caused in the right way" clause? To answer this question, consider what the view would look like without one. Absent this clause, the view is subject to two kinds of counterexamples. The first kind of counterexample stems from the fact that critical reasoning can have side-effects. Consider

> **Side-effects:** You've spent the day engaging in rigorous critical reasoning. You reasoning is so intense it causes you to become dizzy, feel tired, and have a headache.

Though caused by critical reasoning, it would be absurd to hold that your being dizzy, feeling tired, or having a headache is assessable for reasonableness. We need a "caused in the right way" clause to rule out such accidental side-effects.

The second kind of counterexample stems from the fact that critical reasoning can kick off a wayward causal chain that leads you to be in the same state that you would have been in had all gone well. Consider

> **Sugar**: You are engaging in critical reasoning about whether sugar dissolves faster in hot water. You have sufficient evidence that it does, and you are in the process of putting it together. But this process causes you to think back on your chemistry course. As it happens, your professor hypnotized you on the last day of class such that the next time you think of your chemistry course (which is now) you will automatically form the belief that sugar dissolves faster in hot

water. This hypnosis directly causes you to form the belief that sugar dissolves faster in hot water.

Keys: You are engaging in critical reasoning about where you left your keys. You have sufficient evidence that they are under the sofa, and you are in the process of putting it together. But this process causes you to become dizzy. You fall to the ground and find yourself staring at your keys under the sofa. This perception directly causes you to form the belief that your keys are under the sofa.[78]

In both cases, although you are well-positioned to later sustain your beliefs through critical reasoning, at their inception, they are areasonable. Your beliefs in these cases are, in an important respect, like blindsight beliefs. But, unlike blindsight beliefs, critical reasoning was part of the actual causal sequence leading to their formation. The problem lies with the operative mechanism. It was a fluke that critical reasoning caused you to arrive at the correct belief. Your reasoning, to borrow a line from Enç (2003, 105), did what it was supposed to do but not in the way it was supposed to do it. We need a "caused in the right way" clause that rules out these sorts of fortuitous causal paths.

The problems presented by accidental side effects and fortuitous causal paths is striking. It is striking because accounts of proper functioning in biology confront the very same problems.[79] Hearts cause pumping and thumping. But thumping is not the proper function of the heart. Hearts were not selected because they thump. Pumping is what the heart was selected for.[80] Pumping is the heart's job—its telos. Accordingly, pumps are functional effects, while thumps are accidental side-effects. Zebra's stripes are for deterring tsetse flies.[81] But suppose seeing a Zebra's stripes causes you to become dizzy, and this causes you to fall on an enormous pile of tsetse flies—flies that otherwise would have sucked the Zebra's blood. Here the stripes perform their function—they get their job done—but not in the way they were selected to do it. The operative mechanism, stripes causing you to become dizzy and killing tsetse flies, is not the mechanism by which the stripes achieved the deterrence of flies when they were selected for. And hence, this instance of deterrence, because of the fortuitous causal path by which it came about, is not an instance of the stripes

[78] Similar cases of deviance can be found in Wedgwood (2006).

[79] The functional/non-functional effects distinction is a reoccurring theme in the essays in Buller (1999) and Ariew et al. (2002).

[80] Here I am assuming the selected effects theory of functions of the sort pioneered by Millikan (1987) and Neander (1991). But I am open to the generalized version put forward by Garson, which holds that, "A function of a trait is an activity that led to its differential reproduction, or its differential retention, in a population" (2019, 93). This matters for the cases of Swampman and Instant Agents. For, as noted above, at inception, all of their attitudes and actions will be areasonable. But, on the generalized selected effects theory, this need not remain so. As time passes, they can develop etiological functions, and so may become assessable for reasonableness. For discussion of this last point, see Graham (2023).

[81] I borrow this example from Garson (2019), who is drawing on the work of Caro et al. (2014).

performing their function. The goal was achieved and the stripes were involved, but this achievement is not attributable to the stripes doing their job.[82]

As these cases suggest, for effects to qualify as functional effects, we need

> *Functional causation*: X causes Y in the right way for Y to be the functional effect of X if and only if (i) X has the function to Y by mechanism M—X was selected for because it caused Y via M—(ii) X caused Y by M.[83]

Functional causation walls off accidental effects. We can claim that, for any thump, it is not an instance of the heart performing its function. A thump is not a functional effect, because it is not the effect the heart was selected for. Functional causation also walls off the effects of fortuitous causal paths. The deterring of tsetse flies by your crushing them after being made dizzy is not the mechanism by which stripes were selected to deter. So despite this being the effect for which the stripes were selected, this is not an instance of the stripes performing their function. The deterrence was not a functional effect of the stripes.

Does functional causation offer the correct specification of the critical reasoning view's "caused in the right way" clause? If so, that would give us the following precisification: Something is assessable for reasonableness if and only if it is the functional effect of critical reasoning—i.e., it was caused by critical reasoning performing its function by way of the mechanism that was used when it was selected for.[84] Of course, incorporating functional causation into the critical reasoning view requires making good on the idea that critical reasoning has been selected for. But assuming for the moment that we can make good on this idea, the results look promising. If critical reasoning has a function of any sort, it seems safe to assume that causing you to be dizzy, feel tired, or have a headache would qualify as accidental, not functional, effects. We thus can claim that they are not assessable for reasonableness. If critical reasoning has been selected to perform a function by way of a certain mechanism, it seems safe to assume that the mechanism in sugar and keys is not this mechanism. In these cases, the goal was achieved and critical reasoning was involved, but this achievement is not attributable to critical reasoning doing its job.

[82] For further discussion of this point, see Millikan (1984). As she puts it, "Associated with each of the proper functions that an organ or system has is a normal explanation for performance of this function, which tells how that organ or system that species historically managed to perform that function. For example, there are a number of proper functions that *can* be performed by certain systems of the human body, given the presence of appropriate lithium compounds in the bloodstream, but that historically have been performed using calcium. The normal explanations for how these functions are performed make reference to the presence of calcium in the blood rather than lithium" (1984, §1).

[83] For a helpful overview of the kinds of mechanisms I have in mind, see Glennan (2017) and Machamer et al. (2000). On this view, roughly put, "A mechanism for a phenomenon consists of entities (or parts) whose activities and interactions are organized so as to be responsible for the phenomenon" (Glennan 2017, 17).

[84] This appeal to functional effects fits nicely with the lesson from Frankfurt cases. That lesson is driven by the idea that there is a non-accidental, explanatory connection between a particular section of the causal chain and assessability. This same non-accidental, explanatory connection holds between functional effects and their causal chain. For this way of understanding Frankfurt cases, see Heering (2022). For more on the explanatory role of functions, see Garson (2016, §3.1).

Thus, the beliefs in these cases are not functional effects of critical reasoning, and so not assessable for reasonableness.[85]

But is it plausible to claim that critical reasoning has a function? Now it seems fairly obvious that reasoning (or inference), understood broadly, is fitness enhancing. As Quine quipped, "Creatures inveterately wrong in their inductions have a pathetic but praiseworthy tendency to die before reproducing their kind" (1969, 126). But, even if true, we need a more specific claim. We need to show that critical reasoning—not all forms of reasoning or inference—was selected for.

What fitness enhancing effects might reasoning which involves representing support relations provide? The tentative answer I want to suggest is that critical reasoning puts you in a position to support your position.[86] Arriving at new attitudes while recognizing the support relations that got you there puts you in a position to convince others and reconvince yourself. You can provide, should the situation call for it, evidence, reasons, warrant, justification, and so on. You are not blind to why you think and do what you think and do. So you can defend your thoughts and actions to yourself (especially in the future when additional will-power is call for) and to others (especially to distrusting strangers who might otherwise harm you).[87] And being in that position makes you far more likely to survive and reproduce.

Put differently, critical reasoning makes what supports your actions and attitudes sharable across time and people.[88] Being able to share what support your actions and attitudes across time and people enhances cooperation. It gives you means by which you can fight without resorting to violence or fraud. Here is Mercier and Sperber making roughly the same point:

[85] I am not the first to notice this connection between deviant causation and functions. For an account of deviance in the context of intentional action that is very close to the one defended here, see Enç (2003, chp. 4; 2004). For an account of deviance in the context of perception that is very close to the one defended here, see Davies (1983). The account I develop here owes much to Enç's and Davies's discussions.

[86] To be clear, I am not committed to this precise answer. In the end, the details are best left to those working in biology, ethology, and psychology. For an alternative account that focuses on maintaining a shared outlook and aligning our intentions, see Norman (2016). In any case, for the argument to go through, we only need the claim that critical reasoning has a function. And either account supports that contention.

[87] If this account of functions is correct, then critical reasoning, when caused in the right way, will deliver something very close to what Gerken (2012) calls discursive justification: the ability to articulate one's reasons for holding a belief.

[88] If the function of critical reasoning is to put you in a position to support your position to your future-self and others, then there does seem to be something to the idea that the premise-attitudes and conclusion-attitude of critical reasoning are exclusively descriptive. For it appears that convincing by argumentation can only be carried out by citing presumed truths. That is, you convince by citing the content of your descriptive attitudes. You don't cite the content of your directive attitudes. Although you might cite the putative fact that, say, you or others have certain desires, that would still conform to the claim made by premise 3. Mill's proof for the principle of utility (2003, U 4.3) is a good example of someone trying to convince by citing presumed truths about directive attitudes. But the argument is still descriptive through and through.

> Humans differ from other animals not only in their hyperdeveloped cognitive capacities but also, and crucially, in how and how much they cooperate. They cooperate not only with kin but also with strangers; not only in here-and-now ventures but also in the pursuit of long-term goals; not only in a small repertoire of species-typical forms of joint action but also in jointly setting up new forms of cooperation. Such cooperation poses unique problems of coordination and trust. […Reason provides] tools for the kind of rich and versatile coordination that human cooperation requires. By giving reasons in order to explain and justify themselves, people indicate what motivates and, in their eyes, justifies their ideas and their actions. In so doing, they let others know what to expect of them and implicitly indicate what they expect of others. Evaluating the reasons of others is uniquely relevant in deciding whom to trust and how to achieve coordination. […] Reason produces reasons that communicators use as arguments to persuade a reticent audience. Reason, by the same token, helps a cautious audience evaluate these reasons, accept good arguments, and reject bad ones. (2017, 8–9)

And here is Tomasello:

> We must single out, finally, a very special discourse context in human communication with world-changing implications for the process of human thinking: shared decision making. Prototypically, we may imagine as an example collaborative partners—or even a council of elders—attempting to choose a course of action, given that they know together in common ground that multiple courses of action are possible. Given their equal power in their interdependent situation, they cannot just tell the other or others what to do; rather, they must suggest a possible course of action and back it up with reasons. […] With modern humans and their skills of conventional linguistic communication, we get to full-blooded reasoning, where "reasoning" means not just to think about something but to explicate in conventional form—for others or oneself—the reasons why one is thinking what one is thinking. […] Such cooperative argumentation, as we may call it, may be modeled in game theory as a battle of the sexes: our highest goals are collaborative—we will hunt together under all circumstances because otherwise there is zero hope of success—but within that cooperative framework we each argue our case. Critically, in this context, neither of us wants to convince the other if we are in fact wrong about the location of antelopes; each would rather lose the argument and eat tonight than win the argument and go hungry. And so a key dimension of our cooperativeness is that we both have agreed ahead of time, implicitly, that we will go in the direction for which there are the "best" reasons. That is what being reasonable is all about. (2014, 109–110)

In light of the profound benefits that critical reasoning would bestow, it seems highly likely that it was selected for.[89]

[89] For further discussion, see Dennett (1996).

The next issue concerns the mechanism. If critical reasoning's function is to put one in a position to support one's position, what was the mechanism by which it achieved this when it was selected for? One plausible conjecture is that critical reasoning operates by rules built into one's cognitive architecture.[90] This conjecture fits nicely with a computationalist picture of the mind.[91] In addition, it helps explain the popularity of rules in the reasoning literature.[92] However, since the argument does not require that this is correct, I won't try to defend this conjecture. If it is granted that critical reasoning was selected to perform a function, it is safe to assume that there was some mechanism or other by which this function was carried out. And any plausible account of what this mechanism is will rule out the mechanisms by which you arrived at your beliefs in Sugar and Keys.

Let's take stock. We started this section with a problem left over from our attempt to answer the possession question. We needed an account of what it takes to be caused in the right way. We then noted that the kinds of cases that motivated the critical reasoning view's "caused in the right way" clause closely paralleled the kinds of cases that motivated functional causation. This took us to the idea that this was in fact the same problem. The kind of causal chain required for functional effects just is the kind of causal chain required for something to be caused in the right way. If critical reasoning had a function, that would make this idea irresistible. And, as it turns out, it is plausible to hold that critical reasoning has a function. Reasoning that involves recognizing support-relations has the function of putting you in a position to support your position. We thus seem vindicated in incorporating functional causation into the critical reasoning view.

If this much is correct, we have what we need to answer the explanatory question. When a thing has been assigned a job, an end, a goal, and goes to do that job, realize that end, achieve that goal, we can ask how well the job was done, the end was realized, the goal achieved. If your job as an archer is to hit the bullseye and you do that job, we can ask how well your shot (the output) did relative to the job assigned (hitting the bullseye). How close did you get to realizing the end or achieving the goal? The standards of archery come to bear on the output of your performance. Notice that they do not come to bear on the accidental side-effects of your performance. Nor do they come to bear on an arrow that hits the bullseye but is not attributable to you doing your job. If, for example, while shooting you trip, release your hand, and a gust of wind redirects the accidentally fired arrow straight into the bullseye, this result is not assessable. This sort of "shot" is not within the domain of archery. Similarly, if a toaster's job is to make toast, and it does that job, the standards relevant to toasting come to bear on the output of the toaster's performance.[93] But they do not come to bear on the accidental side-effects of the toaster doing its job. Nor do they come to bear on toast caused by the toaster, but not attributable to the toaster doing its job. Such outputs are non-functional effects; they are not the outputs of the toaster doing its job as toaster.

[90] I borrow this way of putting it from Quilty-Dunn and Mandelbaum (2018).

[91] For an overview, see the essays in Haugeland (1981). For a powerful recent defense, see Piccinini (2010, 2020).

[92] See, for example, Boghossian (2014, 2018), Broome (2013, 2019), Quilty-Dunn and Mandelbaum (2018), and McHugh and Way (2018). For doubts about rules, see Warren (2022).

[93] For a related discussion, see Thomson (1997, §6; 2008, chp. 12).

These everyday examples give a feel of the connection between having a telos and being assessable in terms of how well that telos has been achieved. Nevertheless, it might be objected that if toasters and archers have been "selected for," they are not selected for in the same sense as the sense in which critical reasoning has been selected for.

I am sympathetic to this worry.[94] So I suggest that we treat the above examples as merely illustrative. Fortunately, even if we leave aside all cases involving artificially assigned functions, the connection between functional effects and assessability is still readily apparent. If something's job is to convert light into information and it does that job in the way it was selected to do it, we can ask how well this job was performed by the standards relevant to converting light into information. The outputs are assessable by the standards of vision. If something's job is to make skin blend into the immediate environment and does that job in the way it was selected to do it, we can ask how well this job was performed by the standards relevant to blending skin into the immediate environment. The outputs are assessable by the standards of camouflage. Again, the same remarks apply concerning accidental effects and fortuitous causal paths. Accidental side-effects of the visual system doing its job are not assessable by the standards of vision. And if the process of camouflaging itself against a sea of crimson leaves happens, by some fluke, to cause a dramatic rise in body temperature, and this dramatic rise causes the organism's skin to turn the very same shade of crimson, this effect is not subject to the standards of camouflage. For this match is a coincidence, not a functional effect.

The connection between critical reasoning and assessability for reasonableness thus turns out to be an instance of the familiar connection between having a function (having a telos) and being subject to the standards relevant to this function (how well the telos was achieved). Assessability for reasonableness is a specific instance of

> *Functional assessability*: If *Y* is the functional effect of *X*, then *Y* is assessable by the standards relevant to *X*'s function.[95]

If critical reasoning's job is to put one in a position to support one's position and does that job in the way it was selected to do it, we can assess its functional effects by the standards relevant to critical reasoning's function. The outputs are assessable by the standards of reason.[96]

[94] For discussion, see Millikan (1999).

[95] I take this to be an uncontroversial claim. As Burge notes, "Where there are functions, it is apriori that there are standards for fulfilling them" (2010a, 338). For further discussion, see Graham (2023, 251; 2019, 111–112) and Burge (2010a, 338–339; 2020, 40–46).

[96] My aim in this essay is to defend a particular theory of assessability for reasonableness. Yet, the theory defended appears to come with substantive commitments. The standard set by the function of critical reasoning seems to answer the substantive question of what makes something reasonable (or unreasonable). But this is not as straightforward as it might appear. Notice, for example, that there are several different ways that things with functions can fail. They might fail to function property. They might function properly, but be in an abnormal or inhospitable environment such that they fall short of completely fulfilling their function. For more on these failures, see Matthewson and Griffiths (2017). Critical reasoning can fail in all of these ways. So, although I do believe there is a link between functional success (or failure) and being reasonable (or unreasonable), I will not attempt to adjudicate how that link should be understood.

The biconditional posited by the critical reasoning view is thus explained by the more general phenomenon of functional assessability. Things that are caused in the right way by critical reason are assessable for reasonableness, because such things are the functional effects of critical reasoning. They thus are subject to the standards relevant to how well critical reasoning did the job, realized the end, achieve the goal that it was selected to do, realize, achieve.

4 Conclusion

Normativity is often thought dark and mysterious. What I hope to have shown in defending the critical reasoning view is that the domain of reason is no more dark or mysterious than the domain of vision, camouflage, or intelligence. All it takes to be a member is the right causal path. Something is assessable for reasonableness if and only if and because it is the functional effect of critical reasoning.

Acknowledgements I would like to thank Winnie Sung, Douglas Portmore, Nikolaj Jang Lee Linding Pedersen, and Aldrin Relador for helpful comments on earlier drafts. I would also like to thank several anonymous reviewers for the careful attention they gave the argument.

Funding This research was supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 1 (RG125/23).

Declarations

Ethical declarations Not applicable.

References

Anscombe, G. E. M. (1963). *Intention* (2nd ed.). Harvard University Press.
Archer, A. (2015). Reconceiving direction of fit. *Thought: A Journal of Philosophy, 4*(3), 171–180.
Ariew, A., Cummins, R., & Perlman, M. (Eds.). (2002). *Functions: New essays in philosophy of psychology and biology*. Oxford University Press.
Aristotle. (1984). *The Complete works of Aristotle: The revised Oxford translation*. In J. Barnes (Ed.). 2 vols. Princeton University Press.
Berker, S. (2013). Epistemic teleology and the separateness of propositions. *Philosophical Review, 122*(3), 337–393.
Block, N. (2002). Some concepts of consciousness. In D. J. Chalmers (Ed.), *Philosophy of mind: Classical and contemporary readings* (pp. 206–218). Oxford University Press.
Block, N. (2023). *The border between seeing and thinking*. Oxford University Press.
Boghossian, P. (2014). What is inference? *Philosophical Studies, 169*(1), 1–18.

Boghossian, P. (2018). Delimiting the boundaries of inference. *Philosophical Issues, 28*(1), 55–69.
BonJour, L. (1985). *The structure of empirical knowledge*. Harvard University Press.
Broad, C. D. (1954). Emotion and sentiment. *The Journal of Aesthetics and Art Criticism, 13*(2), 203–214.
Broome, J. (2013). *Rationality through reasoning*. Wiley Blackwell.
Broome, J. (2019). A linking belief is not essential for reasoning. In M. B. Jackson & B. B. Jackson (Eds.), *Reasoning: New essays on theoretical and practical thinking* (pp. 32–43). Oxford University Press.
Buller, D. J. (Ed.). (1999). *Function, selection, and design*. State University of New York Press.
Burge, T. (1996). Our entitlement to self-knowledge. *Proceedings of the Aristotelian Society, 96*, 91–116.
Burge, T. (2010a). *Origins of objectivity*. Oxford University Press.
Burge, T. (2010b). Steps toward origins of propositional thought. *Disputatio, 4*(29), 39–67.
Burge, T. (2020). Entitlement: The basis for empirical epistemic warrant. In P. J. Graham & N. J. L. L. Pedersen (Eds.), *Epistemic entitlement* (pp. 37–142). Oxford University Press.
Buss, S. (1999). What practical reasoning must be if we act for our own reasons. *Australasian Journal of Philosophy, 77*(4), 399–421.
Caro, T., Izzo, A., Reiner, R. C., Walker, H., & Stankowich, T. (2014). The function of zebra stripes. *Nature Communications, 5*(1), 1–10.
Chisholm, R. M. (1989). *Theory of knowledge* (3rd ed.). Prentice Hall.
Copp, D. (1997). Defending the principle of alternate possibilities: Blameworthiness and moral responsibility. *Nous, 31*(4), 441–456.
Dancy, J. (1993). *Moral reasons*. Blackwell.
Darwin, C. (2009[1871]). *The descent of man and selection in relation to sex* (Vol. 1). Cambridge University Press.
Davidson, D. (1980). Actions, reasons, and causes. In D. Davidson (Ed.), *Essays on actions and events* (pp. 3–19). Oxford University Press.
Davidson, D. (1987). On knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association, 60*, 441–458.
Davies, M. (1983). Function in perception. *Australasian Journal of Philosophy, 61*(4), 409–426.
Dennett, D. (1984). *Elbow room: The varieties of free will worth wanting*. Oxford University Press.
Dennett, D. (1996). *Kinds of minds: Toward an understanding of consciousness*. Basic Books.
Dorsey, D. (2013). Consequentialism, cognitive limitations, and moral theory. In M. Timmons (Ed.), *Oxford studies in normative ethics* (pp. 180–202). Oxford University Press.
Dreier, J. (1990). Internalism and speaker relativism. *Ethics, 101*(1), 6–26.
Dretske, F. (1993). Can intelligence be artificial? *Philosophical Studies, 71*(2), 201–216.
Drucker, D. (2022). Reasoning beyond belief acquisition. *Noûs, 56*(2), 416–442.
Enç, B. (2003). *How we act: Causes, reasons, and intentions*. Oxford University Press.
Enç, B. (2004). Causal theories of intentional behavior and wayward causal chains. *Behavior and Philosophy, 32*(1), 149–166.
Fischer, J. M. (2003). 'Ought-implies-can', causal determinism and moral responsibility. *Analysis, 63*(3), 244–250.
Fischer, J. M. (2012). *Deep control: Essays on free will and value*. Oxford University Press.
Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.
Fish, W. (2010). *Philosophy of perception: A contemporary introduction*. Routledge.
Fletcher, G., & Ridge, M. (Eds.). (2014). *Having it both ways: Hybrid theories and modern metaethics*. Oxford University Press.
Foley, R. (1991). Evidence and reasons for belief. *Analysis, 51*(2), 98–102.
Foot, P. (1978). *Virtues and vices*. Oxford University Press.
Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *The Journal of Philosophy, 66*(23), 829–839.
Frost, K. (2014). On the very idea of direction of fit. *Philosophical Review, 123*(4), 429–484.
Garson, J. (2016). *A critical overview of biological functions*. Springer.
Garson, J. (2019). *What biological functions are and why they matter*. Cambridge University Press.
Gerken, M. (2012). Discursive justification and skepticism. *Synthese, 189*(2), 373–394.
Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology, 65*(2), 221–233.
Glennan, S. (2017). *The new mechanical philosophy*. Oxford University Press.

Graham, P. J. (2019). Why is warrant normative? *Philosophical Issues, 29*(1), 110–128.
Graham, P. J. (2023). Proper functionalism and the organizational theory of functions. In L. R. G. Oliveira (Ed.), *Externalism about knowledge* (pp. 249–276). Oxford University Press.
Gregory, A. (2019). Why do desires rationalize actions? *Ergo, 5*(40), 1061–1081.
Gregory, A. (2021). *Desire as belief: A study of desire, motivation, and rationality*. Oxford University Press.
Grice, H. P. (1971). Intention and uncertainty. *Proceedings of the British Academy, 57*, 263–279.
Haji, I. (2002). *Deontic morality and control*. Cambridge University Press.
Haji, I. (2023). Obligation and Responsibility. Oxford: Oxford University Press.
Haugeland, J. (Ed.). (1981). *Mind design*. MIT Press.
Heering, D. (2022). Actual sequences, Frankfurt-cases, and non-accidentality. *Inquiry, 65*(10), 1269–1288.
Hieronymi, P. (2009). Believing at will. *Canadian Journal of Philosophy, 39*, 149–187.
Hume, D. (1960) [1739]. *A treatise of human nature: Being an attempt to introduce the experimental method of reasoning into moral subjects*. In L. A. Selby-Bigge (Ed.). Oxford University Press.
Jackson, F. (1991). Decision-theoretic consequentialism and the nearest and dearest objection. *Ethics, 101*(3), 461–482.
Joyce, J. M., & Weatherson, B. (2019). Accuracy and the imps. *Logos & Episteme, 10*(3), 263–282.
Kagan, S. (2002). Kantianism for consequentialists. In A. W. Wood (Ed.), *Groundwork for the Metaphysics of Morals* (pp. 111–157). Yale University Press.
Kant, I. (2002). *Groundwork for the metaphysics of morals*. In A. W. Wood, J. B. Schneewind (Eds.). Yale University Press.
Kavka, G. S. (1983). The toxin puzzle. *Analysis, 43*(1), 33–36.
Korsgaard, C. M. (2009). *Self-constitution: Agency, identity, and integrity*. Oxford University Press.
Kriegel, U. (2022). Moral judgment and the content-attitude distinction. *Philosophical Studies, 179*(4), 1135–1152.
Lehrer, K. (1990). *Theory of knowledge*. Routledge.
Little, M. O. (1997). Virtue as knowledge: Objections from the philosophy of mind. *Noûs, 31*(1), 59–79.
Lord, E. (2015). Acting for the right reasons, abilities, and obligation. In R. Shafer-Landau (Ed.), *Oxford studies in metaethics* (pp. 26–52). Oxford University Press.
Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25.
Marušić, B., & Schwenkler, J. (2018). Intending is believing: A defense of strong cognitivism. *Analytic Philosophy, 59*(3), 309–340.
Matthewson, J., & Griffiths, P. E. (2017). Biological Criteria of Disease: Four Ways of Going Wrong. *Journal of Medicine and Philosophy, 42*(4), 447–466.
McCormick, M. (2014). *Believing against the evidence: Agency and the ethics of belief*. Routledge.
McHugh, C. (2013). Epistemic responsibility and doxastic agency. *Philosophical Issues, 23*, 132–157.
McHugh, C. (2017). Attitudinal control. *Synthese, 194*(8), 2745–2762.
McHugh, C., & Way, J. (2018). What is reasoning? *Mind, 127*(505), 167–196.
McKenna, M. (2008). Frankfurt's argument against alternative possibilities: Looking beyond the examples. *Noûs, 42*(4), 770–793.
McLaughlin, P. (2007). On selection of, for, with, and against. In P. Machamer & G. Wolters (Eds.), *Thinking about cause: From Greek philosophy to modern physics* (pp. 265–283). University of Pittsburgh Press.
Mele, A. R. (2003). *Motivation and agency*. Oxford University Press.
Mele, A. R. (2006). *Free will and luck*. Oxford University Press.
Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.
Mill, J. S. (2003) [1861/1859]. *Utilitarianism and on liberty: Including Mill's 'Essay on Bentham' and selections from the writings of Jeremy Bentham and John Austin* (2nd ed.). In M. Warnock (Ed.). Blackwell Publishing.
Millikan, R. G. (1984). Naturalist reflections on knowledge. *Pacific Philosophical Quarterly, 65*(4), 315–334.
Millikan, R. G. (1987). *Language, thought, and other biological categories: New foundations for realism*. MIT press.
Millikan, R. G. (1995). Pushmi-pullyu representations. *Philosophical Perspectives, 9*, 185–200.
Millikan, R. G. (1999). Wings, spoons, pills, and quills: A pluralist theory of function. *The Journal of Philosophy, 96*(4), 191–206.

Millikan, R. G. (2004). *Varieties of meaning*. MIT Press.
Millikan, R. G. (2017). *Beyond concepts: Unicepts, language, and natural information*. Oxford University Press.
Nagel, T. (1970). *The possibility of altruism*. Princeton University Press.
Neander, K. (1991). Functions as selected effects: The conceptual analyst's defense. *Philosophy of Science, 58*(2), 168–184.
Neander, K. (2009). Teleological theories of mental content. In M. Ruse (Ed.), *The oxford handbook of philosophy of biology* (pp. 381–409). Oxford University Press.
Neta, R. (2013). What is an inference? *Philosophical Issues, 23*, 388–407.
Nichols, S., & Stich, S. P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds*. Oxford: Oxford University Press.
Norman, A. (2016). Why we reason: Intention-alignment and the genesis of human rationality. *Biology & Philosophy, 31*, 685–704.
Nussbaum, M. C. (2001). *Upheavals of thought: The intelligence of emotion*. Cambridge University Press.
Parfit, D. (1997). Reasons and motivation. *Proceedings of the Aristotelian Society, supplementary volumes, 71*, 99–130.
Parfit, D. (2011). *On what matters* (Vol. 1). In S. Scheffler (ed.). Oxford University Press.
Pettit, P. (2016). Broome on reasoning and rule-following. *Philosophical Studies, 173*(12), 3373–3384.
Piccinini, G. (2010). The mind as neural software? Understanding functionalism, computationalism, and computational functionalism. *Philosophy and Phenomenological Research, 81*(2), 269–311.
Piccinini, G. (2020). *Neurocognitive mechanisms: Explaining biological cognition*. Oxford University Press.
Platts, M. (1997). *Ways of meaning: An introduction to philosophy of language* (2nd ed.). MIT Press.
Portmore, D. W. (2019). *Opting for the best: Oughts and options*. Oxford University Press.
Price, H. (1989). Defending desire-as-belief. *Mind, 98*(389), 119–127.
Quilty-Dunn, J., & Mandelbaum, E. (2018). Inferential transitions. *Australasian Journal of Philosophy, 96*(3), 532–547.
Quine, W. V. O. (1969). *Ontological relativity*. Columbia University Press.
Raz, J. (1999). *Engaging reason: On the theory of value and action*. Oxford University Press.
Regan, D. H. (2003). How to be a Moorean. *Ethics, 113*(3), 651–677.
Rescher, N. (1988). *Rationality: A philosophical inquiry into the nature and the rationale of reason*. Oxford University Press.
Ross, W. D. (1939). *Foundations of Ethics*. Oxford Oxford University Press.
Sartorio, C. (2016). *Causation and free will*. Oxford University Press.
Sartorio, C. (2017). Actual causes and free will. *Disputatio, 9*(45), 147–165.
Scanlon, T. M. (1998). *What we owe to each other*. Harvard University Press.
Scanlon, T. M. (2007). Structural irrationality. In R. E. Goodin, G. Brennan, F. Jackson, & M. Smith (Eds.), *Common minds: Themes from the philosophy of Philip Pettit* (pp. 84–103). Oxford University Press.
Scheffler, S. (2011). Valuing. In R. J. Wallace, R. Kumar, & S. Freeman (Eds.), *Reasons and recognition: Essays on the philosophy of T.M. Scanlon* (pp. 23–42). Oxford University Press.
Schiffer, S. R. (1981). Truth and the theory of content. In H. Parret & J. Bouveresse (Eds.), *Meaning and understanding* (pp. 204–222). de Gruyter.
Schueler, G. F. (1995). *Desire*. MIT Press.
Searle, J. R. (2001). *Rationality in action*. MIT Press.
Setiya, K. (2008). Believing at will. *Midwest Studies in Philosophy, 32*, 36–52.
Shah, N. (2006). A new argument for evidentialism. *The Philosophical Quarterly, 56*(225), 481–498.
Shah, N. (2008). How action governs intention. *Philosopher's Imprint, 8*(5), 1–19.
Shea, N. (2014). Reward prediction error signals are meta-representational. *Noûs, 48*(2), 314–341.
Sidgwick, H. (1907). *The methods of ethics* (7th ed.). Macmillan.
Sinhababu, N. (2017). *Humean nature: How desire explains action, thought, and feeling*. Oxford University Press.
Smith, M. (1992). Valuing: Desiring or believing? In D. Charles & K. Lennon (Eds.), *Reduction, explanation, and realism* (pp. 323–359). Oxford University Press.
Smith, M. (1994). *The moral problem*. Blackwell.

Smith, M. (2003). Rational capacities, or: How to distinguish recklessness, weakness, and compulsion. In S. Stroud & C. Tappolet (Eds.), *Weakness of will and practical irrationality* (pp. 17–38). Oxford University Press.
Smith, M. (2012). Four objections to the standard story of action (and four replies). *Philosophical Issues, 22*(1), 387–401.
Smith, M. (2013). The ideal of orthonomous action, or the how and why of buck-passing. In D. Bakhurst, M. O. Little, & B. Hooker (Eds.), *Thinking about reasons: Themes from the philosophy of Jonathan Dancy* (pp. 51–74). Oxford University Press.
Staffel, J. (2019). Attitudes in active reasoning. In M. B. Jackson & B. B. Jackson (Eds.), *Reasoning: New essays on theoretical and practical thinking* (pp. 44–67). Oxford University Press.
Tenenbaum, S. (2021). The guise of the good. In R. Chang & K. Sylvan (Eds.), *The Routledge handbook of practical reason* (pp. 226–236). Routledge.
Thomson, J. J. (1997). The right and the good. *The Journal of Philosophy, 94*(6), 273–298.
Thomson, J. J. (2008). *Normativity*. Open Court.
Timmerman, T., & Cohen, Y. (2016). Moral obligations: Actualist, possibilist, or hybridist? *Australasian Journal of Philosophy, 94*(4), 672–686.
Tomasello, M. (2014). *A natural history of human thinking*. Harvard University Press.
Tomasello, M. (2022). *The evolution of agency: Behavioral organization from lizards to humans*. MIT Press.
Wallace, J. R. (1990). How to argue about practical reason. *Mind, 99*(395), 355–385.
Warren, J. (2022). Functionalism about inference. *Inquiry*, 1–25.
Watson, G. (1975). Free agency. *The Journal of Philosophy, 72*(8), 205–220.
Way, J., & Whiting, D. (2017). Perspectivism and the argument from guidance. *Ethical Theory and Moral Practice, 20*(2), 361–374.
Weatherson, B. (2008). Deontology and descartes's demon. *The Journal of Philosophy, 105*(9), 540–569.
Wedgwood, R. (2006). The normative force of reasoning. *Noûs, 40*(4), 660–686.
Wedgwood, R. (2017). *The value of rationality*. Oxford University Press.
Widerker, D. (1991). Frankfurt on 'ought implies can' and alternative possibilities. *Analysis, 51*(4), 222–224.
Widerker, D., & McKenna, M. (2006). Introduction. In D. Widerker & M. McKenna (Eds.), *Moral responsibility and alternative possibilities: Essays on the importance of alternative possibilities* (pp. 1–16). Ashgate Press.
Williams, B. (1973). A critique of utilitarianism. In *Utilitarianism: For and against* (pp. 3–67). Cambridge University Press.
Williams, B. (1981). *Moral luck: Philosophical papers, 1973–1980*. Cambridge University Press.
Williamson, T. (2000). *Knowledge and its limits*. Oxford University Press.