From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Philosophy

2021

The Publicity of Practical Reasons

Ayars, Alisabeth

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp01s4655k735

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

THE PUBLICITY OF PRACTICAL REASONS

Alisabeth Ayars

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE BY

THE DEPARTMENT OF PHILOSOPHY

Adviser: Gideon Rosen

Adviser: Johann Frick

Adviser: Michael Smith

September 2021

© Copyright by Alisabeth Ayars, 2021. All rights reserved.

Abstract

This dissertation consists of three papers relating to the theme of the shared nature of practical reasons. The first two, taken together, aim to derive and explain the thesis that practical reasons are public: If R is a reason for X to do A in C, then R is a reason for anyone not to interfere with X's doing A in C. I derive this significant thesis about reasons from an independently motivated non-cognitivist account of normative judgment, in particular, "ought" judgments and reasons judgments. The third paper provides a theory of blameworthiness which grounds the blameworthiness of agents in failure to recognize and respond to these public reasons, both practical and theoretical.

Acknowledgements

First and foremost, I am incredibly grateful to my main advisers, Gideon Rosen and Johann Frick, for the support, attention, and philosophical insight they have provided me over my years at Princeton.

My frequent conversations with Gideon have been essential to the development of all the ideas in this dissertation. At every point, Gideon has been there to provide insight when I was lost, and to help me turn vague thoughts into fully formed claims and arguments. Gideon has been a tremendous mentor to me, and I am deeply grateful.

Johann Frick has also been a wonderful mentor. In my third year, I asked Johann to co-advise me — one of the best decisions I made in grad school! Johann's consistent and sharp philosophical feedback has been essential to my philosophical progress and has greatly improved this dissertation. I'll also be forever grateful for his encouragement and support on the job market this year, which helped me get through many low points.

I thank Michael Smith for the many conversations we had at the beginning stages of this dissertation. At that time, the ideas were very fuzzy. Michael provided insights and clarifications that were critical to their development.

I also thank Tom Kelly and Elizabeth Harman. My work has been much enhanced by responding to Tom's (often relentless) objections. His practical advice has also proved invaluable. Liz has been very supportive to me at Princeton, especially through the UCHV Graduate Prize Fellow Seminar, and also provided excellent philosophical feedback, especially on the third paper.

I am grateful to Sarah McGrath for her feedback in first-year seminar and dissertation seminar, and for her metaethics course last spring, which I found extremely illuminating. It provided a critical high point for me just as coronavirus was making everything terrible.

I thank Sam Berstler, for her wisdom, support, and philosophical insight over the past two years. I'm also grateful to Ellie Cohen, who has taught me many things, including much of what I know about logic. Finally, I thank Philip Pettit and Harvey Lederman for various things they have taught me and the feedback they have given me on ideas.

Table of Contents

Abstract.....iii

Introduction.....1

First Paper: “Deciding for Others: An Expressivist Theory of Normative Judgment”.....13

Second Paper: “An Explanation for the Essential Publicity of Practical Reasons”.....43

Third Paper: “Blaming for Unreasonableness: Accountability Without Ill Will”.....78

Introduction

I. The Publicity Thesis

Suppose I am stepping on your broken toe, causing you pain. As Nagel observes in *The Possibility of Altruism* (1970), you would not merely dislike this; you would resent it. Implicit in your resentment is the supposition that I should stop, and hence, that I have a reason to stop. If this resentment is warranted, it follows that your pain has a dual significance. It gives you a reason to pull your foot away, and it also gives me a reason to (at least) not interfere with your pulling your foot away. Your reason is also a reason for me.

Consider also an example from Korsgaard (2009). Suppose I'm a teacher and you're a student, and we need to schedule an appointment. I suggest meeting after class, thinking this will be convenient for you, but it turns out it's not. Once I learn this, I automatically have a reason to suggest a different time, even if I could coerce you into meeting now, which is what would be in my narrow self-interest. The inference I draw is reflexive: "The time I suggested isn't good for you, and therefore it isn't good for us, and it follows from that that it isn't after all good for me, and so I need to suggest another time" (2009: 192). Your reasons are reasons for me. It is not just that your reasons are relevant to what I should do, because, e.g., they will help me predict your behavior or what you will agree to. Rather, they bear on what I ought to do simply because they are your reasons.

This seems to hold in full generality: a reason for you is always a reason for me. Even in competitive contexts, the reasons of other people plausibly give us reasons. If you and I are shipwrecked, fighting for the last plank, it may be that each of us should, all things considered, try to get the plank for herself. But even in this context, your reasons still give me reasons. The fact that you need the plank to save your life gives me some reason to let you have it. Reasons have an intersubjective or shared/public dimension: they create reasons for others.

Moreover, others' reasons seem to give us reasons directly. It does not matter whether it is in my self-interest to respond to your reasons, or whether I happen to care about you. I

am rationally required to respond to them simply because they are your reasons. As Nagel put it: other-regarding reasons do "not depend on intermediate factors such as one's own interests or one's antecedent sentiments of sympathy or benevolence" (Nagel 1970: 15-16).

Nagel famously argued that every agent-relative reason implies an agent-neutral one — every reason for X to Φ is a reason for everyone to see to it that X Φs — and moreover, a reason of equal force. So, for instance, if you have a reason to climb Mt. Kilimanjaro because it is your lifelong goal, this fact is a reason of equal force for me to promote your climbing Kilimanjaro. Nagel argued that embracing an impersonal perspective of this sort is necessary to avoid solipsism. Unless we take an impersonal view — one in which our judgments about the reasons to promote someone's Φing are not affected by the standpoint from which these reasons are judged — we fall into a kind of practical solipsism amounting to the denial of the fact that we are just one person among many.

Nagel has since rejected this view ; indeed, it is implausible. While your reason to climb Kilimanjaro may have some normative significance for me, it is not plausible that I have just the same reason to promote your climbing it as you do. The fact that it is your project to climb Kilimanjaro, not mine, makes some difference.

Some philosophers now defend a more moderate but still significant claim about reasons (Wallace 2009, Darwall 1983, Korsgaard 1996, Bertea 2017). They contend that reasons are "public": their normative force necessarily extends "across the boundaries between people" (Korsgaard 2009: 191). But they deny that a reason for X to Φ is always a reason to promote X's Φing.

According to Wallace, the significance of the reasons of other people is that they always generate reasons for others to not interfere with the agent's Φing. The Nagelian thesis, according to Wallace, assumes a teleological conception of practical normativity according to which all reasons are derived from the value of states of affairs. This is a conception of normativity that Wallace rejects, as does Korsgaard (1993). But, Wallace suggests, we can divorce the publicity thesis from this conception, by identifying the significance of others' reasons to be the pro tanto requirement of interference rather than promotion:

> Here is one possibility that is opened up by the non-teleological conception of practical justification. The most basic expression of the normative significance that my reasons have for others is the (pro tanto) requirement of non-interference. Thus if my friendship with Peter is a genuine reason for me to devote special attention to Peter's needs, then other agents have reason not to interfere with the activities whereby I go about attending to those needs. Furthermore, the same considerations that provide my special reasons for engaging in these activities also provide the reasons why others should not interfere with them. The fact that my X-ing would help my friend Peter, who is in a bad way, gives me reason to do X; but it also gives you reason not to interfere with my X-ing (Wallace 2009: 477)

Call Wallace's thesis the

> **Publicity Thesis:** Necessarily, if R is a reason for X to Φ in C, then R is a reason for anyone to not interfere with X's Φing in C.

This Publicity Thesis is credible — more credible than the strong Nagelian thesis. If you have reason to climb Kilimanjaro, this may not give me reason of equal force to promote your climbing it. Indeed it may give me no reason at all to promote this idiosyncratic project of yours. But plausibly, it gives me some reason to not interfere with your climbing it. The Publicity Thesis does not imply that all reasons are agent-neutral, or that we all have reasons to promote the same ends. It is just the claim that the considerations that recommend certain actions to one person also provide reasons for others not to interfere. In the shipwreck case, your reasons to get the plank for yourself give me some reason to not interfere — even if these reasons are overridden by my reasons to try to get it for myself.

If true, the Publicity Thesis is a deep fact about how reasons work. Practical reasons are never purely private. The alternative view, which certainly seems conceivable, is that while others have reasons of their own, they do not necessarily make any claims on other people. An egoist thinks that other people have reason to benefit themselves but that these reasons do not necessarily provide him with reasons to do anything. It is a striking feature of normative reality that it is not this way. Other people and their reasons always matter for everyone, everywhere.

As noted, the Publicity Thesis has a certain initial plausibility. When I am in position to interfere with you in the pursuit of your rational ends, I am always under some pressure not

to do so. Of course, however, questions will arise. Suppose you have some reason to do something that it would be morally wrong for you to do — rob a bank, for instance. Does it follow that I have some reason not to interfere?

It is clear that I don't have a serious reason — the sort of reason it makes sense to dwell on. But as Schroeder (2007) has emphasized, it can be hard to tell the difference between a very weak reason, or an overridden reason, and no reason at all; so, it can be hard to test claims like the Publicity Thesis by considering hypothetical cases, since apparent counterexamples can always be explained away.

Given this, it would be good to have an argument for the Publicity Thesis — a derivation of the thesis from deeper and independently plausible principles. And of course, even if we find the principle plausible and do not insist on the sort of argument that might persuade a skeptic, we should still be in the market for an explanation for it. So, we ask: What could possibly explain the truth of the Publicity Thesis?

One possibility, of course, is that it is just a brute fact about normative reality that others' reasons always generate pressure to not interfere: an axiom in the theory of practical rationality (analogous to the Principle of Utility for the utilitarian). This would not entail that the principle could not be supported, for example, by appeal to coherence with our other views. But it would entail that there is no explanation for its truth. On this view, the fact every reason for one is a reason for all is simply part of the furniture of the universe.

Wallace is happy to endorse the Publicity Thesis simply on the basis of its reflective support and coherence with our other views. The Publicity Thesis, he contends, follows from "our best substantive understanding of what we (individually and severally) have reason to do" (2009: 471). Evidence for the Publicity Thesis can be found in the first-order normative judgments that we are inclined to endorse in reflection. Our normative judgments follow a pattern that supports the Publicity Thesis: notably, "the judgment that S lacks good reason to do X seems necessary for our concluding that our interfering with S's X-ing is something we may be unproblematically be entitled to do" (485). By contrast, if we believe S to have reason to X, we take ourselves to be under at least some rational pressure to not interfere.

Wallace has us consider the greater pressure we feel against interference when someone is using violence in self-defense (which is a genuine reason for violence) versus when someone is using it to take revenge for an imagined slight (which is not a genuine reason). We only feel unproblematically entitled to interfere in the latter case. For Wallace, the fact that the Publicity Thesis is supported by our reflective judgments suffices for accepting it (484). But no further explanation for its truth is required; indeed there may be no such explanation.

While this reflective support is evidence for the truth of the Publicity Thesis, it would be disappointing if we could find no further reinforcement or explanation for it. As noted earlier, it can be hard to tell the difference between a very weak reason, or an overridden reason, and no reason at all. This makes it hard to test claims like the Publicity Thesis by considering hypothetical cases. Wallace overestimates the clarity of our intuitions across the full range of cases and fails to consider the cases that would provide the most stringent test, e.g., the case in which someone has a reason to act immorally.

Moreover, the Publicity Thesis seems to be the sort of thesis that cries out for further explanation. We may all agree, as a substantive matter, that the reasons of other people bear on what we ought to do. But why should this be the case? Why should one take the reasons of others into account, rather than acting only for one's own sake? As Sidgwick famously observed, the boundaries between individuals are real and substantial, and there is no obvious justification for the claim that we must take others' interests and reasons to have normative force for us:

> It would be contrary to Common Sense to deny that the distinction between any one individual and any other is real and fundamental, and that consequently 'I' am concerned with the quality of my existence as an individual in a sense, fundamentally important, in which I am not concerned with the quality of the existence of other individuals: and this being so, I do not see how it can be proved that this distinction is not to be taken as fundamental in determining the ultimate end of rational action for an individual (Sidgwick 1907, 498).

If there are reasons to care about the interests of others, they are a very different sort than the familiar reasons of self-interest. Why do we have these reasons in addition to reasons of self-interest (or reasons to satisfy our own given desires)? When we ask for an explanation of the

publicity thesis, we are asking a version of the foundational question in moral theory that goes back to Plato: Why do non-self-interested reasons exist, and where do they come from?

Some philosophers argue that the Publicity Thesis does have a deep explanation (Korsgaard 1996, Bertea 2017). Unlike Wallace, they do not think that the Publicity Thesis is a substantive, "synthetic" truth, but rather that it follows from the very nature or concept of a reason — or an account what we are doing when we ascribe reasons to others. According to Korsgaard practical reasons are "public in their very essence" (Korsgaard 1996: 135). We cannot coherently conceive of a reason that binds one person but not others. This sort of view is designed to answer the "why?" question in decisive way, by providing a justification for the Publicity Thesis that should lead any reflective, truly rational skeptic to revise his view, irrespective of his substantive sensibility.

> … responding to another's reasons as normative is the default position—just like hearing another's words as meaningful is the default position. It takes work to ignore someone else's reasons; it's nearly as hard to be bad as it is to be good. And that's because reasons are public (2009: 202).

Korsgaard offers a Wittgensteinian private-language argument for this conclusion which aims to show that we cannot coherently conceive of a private reason. Wittgenstein argued that there must be public rules governing the meanings of words. Otherwise, there would be no way to determine whether a person made a mistake in using a word, and the possibility of error is necessary for a word to have meaning. Korsgaard contends that analogously, there must be public rules governing the ascription of reasons for such ascriptions to be meaningful at all.

Korsgaard's argument, however, is widely rejected. As Gert (2002) points out, even it is true that there must be public rules governing the ascription of reasons, this would not rule out private reasons any more than the existence of public rules governing the use of words rule out indexical terms like 'myself' and 'yours'. The rules governing the ascription of self-interested reasons are public — there is the possibility of making a mistake in ascribing a reason of self-interest to an agent, by failing to conform to these rules — yet this does not show that reasons of self-interest are "public" in the relevant sense (i.e., have normative force for everyone).

Indeed, general skepticism about the claim that reasons are "essentially" public is well-grounded. The skeptic who is happy to speak of reasons, but thinks that your reasons have no claim whatsoever on me, does not seem confused about what reasons are or in violation of formal constraints on the ascription of reasons.

Nonetheless, I think there is reason to think this strong thesis is true. In the first two papers of this dissertation, I offer a new explanation for the Publicity Thesis — one that, if true, entails that reasons are "essentially" public, though in a sense that differs somewhat from Korsgaard's (or Nagel's). The Publicity Thesis, I show, follows from a plausible form of non-cognitivism about normative judgment, and in particular about reasons judgments. If this form of non-cognitivism is true, then anyone who judges that R is a reason for X to Φ is rationally committed to judging that R is a reason to not interfere with X's Φing. Hence, everyone is committed to every instance of the Publicity Thesis by the nature of normative judgment, in particular judgments about reasons and "ought" judgments.

The non-cognitivist view "entails" the Publicity Thesis indirectly — by entailing that no one can coherently reject any instance of it. Reasons are "essentially" public in the sense that their publicity follows from an account of what we are doing when we ascribe reasons to others, rather than any substantive normative commitments. The Publicity Thesis is not a synthetic truth, as Wallace thinks, but rather follows from the nature of normative judgment. Like Korsgaard's view, this explanation for the Publicity Thesis is designed to answer the "why?" question in decisive way — if viable, it is the sort of justification for the Publicity Thesis that should lead any reflective skeptic to revise his view regardless of his substantive sensibility.

Since the derivation of the Publicity Thesis is complicated, it's not surprising that the alternatives — e.g., the position of the egoist who thinks that the reasons of other people make no claim on him — do not strike us as confused. Nonetheless, if the non-cognitivist view is true, they are confused. Since the non-cognitivist view is plausible, we have reason to take this possibility seriously.

Importantly, my non-cognitivist view would explain why the reasons of others provide us with reasons not to interfere without at the same time providing us with reasons to promote

their ends. This is important, since unlike the Publicity Thesis, this strong Nagelian thesis is far from plausible. One outstanding problem in this area is to explain why, having accepted that other people's reasons provide us with reasons not to interfere, we should not go "all the way" and accept the strong view that would give every rational agent the same ends (Wallace 2009). One virtue of the present account is that it explains why it is not "arbitrary" to draw the line where the Publicity Thesis draws it.

II. The Non-Cognitivist View (Papers 1 and 2)

By way of preview, on the non-cognitivist view I defend, the normative judgment that X should Φ consists in a kind of decision that X Φ. When the judgment is first-personal—e.g., my judgment that I should Φ — the view is similar to Gibbard's plan expressivism, though the state I call "decision" differs somewhat from a Gibbard-style plan. (I define a "decision" in functional terms, as a state that is distinct from intention but whose role is to issue in intentions).

The deep difference between the views shows in the account of third-personal judgments. On Gibbard's view (the Subtle View) the judgment that Mary should Φ is a "contingency plan" on the thinker's part to Φ if she finds herself in Mary's exact situation. On my view (the Simple View) the judgment is a decision for Mary that Mary Φ.

> **Simple View:** Just as my first-personal judgment that I should Φ is a decision that I Φ, my third-personal judgment that X should Φ is a decision that X Φ.

The Publicity Thesis follows from the Simple View and a corresponding non-cognitivist view of reasons judgments, as I show in the second paper.

My first paper, "Deciding for Others: An Expressivist Theory of Normative Judgment," focuses on defending the Simple View. (The Simple View is called "Simple Plan Expressivism" in that paper).

The Simple View says that my third-personal judgment is a decision on my part that Mary leave. This is surprising — the notion of deciding for another person seems unfamiliar. But I argue first that the Simple View solves intractable problems for Gibbard's view, and second that the core idea is not as strange as it may seem.

I illustrate the idea using an example. Suppose we're at a restaurant and I ask you to decide what I will order. You deliberate, taking my reasons into account; and then voice your conclusion with an imperative — "Get the burrito"— or a normative assertion: "You should get the burrito". Your judgment aims to settle for me the practical question of what to order and it functions well if it issues in an intention on my part to get the burrito. Hence it is a kind of decision for me. The Simple View has the consequence that the role of third-personal normative judgment is social, aimed at guiding the actions of others.

Another piece of evidence that your judgment deserves to be called a "decision" for me is that certain rational norms characteristic of decisions appear to be operative. If there is only one burrito left and you decide on my behalf that I'm to get the last burrito (i.e., judge that I should get it), you cannot coherently decide to get it yourself (i.e., judge that you should get it). We can explain this by noting that decisions are subject to a co-realizability constraint: it is irrational to decide that A Φ and B Ψ if it is impossible that [A Φ and B Ψ]. I argue that this rational constraint on normative judgment is independently plausible. The defense of this norm is a critical step in the defense of the Publicity Thesis.

Given the co-realizability principle, the Simple View entails that if you judge that X should Φ, you are committed to not interfering with X's Φing. The judgment that X should Φ is a decision that X Φ; the decision to interfere with X's Φing is a decision to do something incompatible with X's Φing. These judgments are non-co-realizable, and therefore incoherent.

So far this is an account of all-things-considered "should" judgments. But the Publicity Thesis is a thesis about reasons. To close the gap, we need an account of judgments about reasons. My second paper, "An Explanation for the Essential Publicity of Practical Reasons," provides such an account and shows how the Publicity Thesis can be derived.

In that paper I adopt the common non-cognitivist view that judgments reasons consist in "weighing considerations in favor" of decisions. To judge that R is a reason to Φ is to be disposed to weigh R in favor of deciding to Φ (Gibbard 2003).

Applying this picture to the Simple View, we say that to judge that R is a reason for Mary to Φ is to be disposed to weigh R in favor of deciding that Mary Φ. Since deciding that Mary Φ commits you to deciding not to interfere with Mary's Φing, weighing R in favor of deciding that Mary Φ commits you to weighing R in favor of deciding against interfering with Mary's Φing.

The intuitive idea can be gleaned by considering the case in which I ask you to decide for me what to get at the restaurant. If you judge that something is a reason for me to get the last burrito, you will — if you are rational — weigh it against deciding to get it yourself. Thus, a reason for you is a reason for me. The argument generalizes, and the Publicity Thesis follows.

The Publicity Thesis is a formal thesis. It says that if you judge that R is a reason for X to Φ, you are committed to weighing R against deciding not to do anything incompatible with X's Φing. But it places no constraints on what R may be. Still, the formal principle is highly significant. It entails that the only possible reasons (or at least, the only reasons we can coherently judge to be reasons) are reasons that we share. My project derives and explains this from an independently motivated account of normative judgment. The fact that Publicity follows from the account serves both as support for the account — since Publicity is plausible — and as an explanation for the Publicity Thesis, a deep fact that we might otherwise be forced to regard as brute.

On the account of normative judgment I defend, what we are doing when we make normative judgments about what others should do, or about their reasons, is deliberating from the perspective of others and ratifying or not courses of action *for them*. Previous attempts to derive an other-regarding commitment from the nature of normative judgment have tended to focus on our disposition to step back from our own impulses and ratify our not the course of action they seem to favor (Korsgaard 1996). They argue that we are (somehow) committed, from within this practical perspective, to regarding others' reasons as normative for us.

In contrast, on the present view, it is our disposition to deliberate from the perspective of other individuals and ratify or not courses of action for them that explains the other-regarding commitment. We cannot take up this perspective without committing ourselves to regarding the reasons of other people as normative for us — since to ascribe a reason to another, from this perspective, already carries with it such a pro tanto commitment to non-interference.

III. Blameworthiness (3rd Paper)

Recall the example from Nagel (1970) described earlier. I am stepping on your broken toe, causing you pain. Nagel says you would not merely dislike this, you would resent it. Implicit in your resentment is the supposition that I should stop, and hence, that I have a reason to stop.

When we blame or resent someone for Φing, implicit in the resentment is the person should have done otherwise, and hence that they had reason not to do otherwise. The justifiability of resentment implies that we have other-regarding obligations, and hence other-regarding reasons. Moreover, we can legitimately resent people even when the action was in the agent's interest, and even when the agent had no desire to do otherwise. Any rational being can resent illegitimate interference from any other. The whole business presupposes that there are always reasons of non-interference of this sort — reasons that do not depend on contingent desires or self-interest. Anyone concerned with the justifiability of blame and punishment thus has an interest in showing how other's reasons give us reasons independently of intermediary factors like self-interest and our antecedent desires.

My account is one possible explanation for how this could be. Having provided this explanation, I turn my attention to blameworthiness. In my third paper, "Blaming for Unreasonableness: Accountability Without Ill Will", I provide a theory of blameworthiness which grounds the blameworthiness of agents in failure to recognize and respond to public reasons, both practical and theoretical.

The paper argues against the following quality of will condition, endorsed by many theorists of moral responsibility: X is blameworthy for a wrong act A only if in doing A, X expressed ill will. I argue that the quality of will condition is in fact false to ordinary moral practice, because ordinary practice licenses blame for agents who act wrongly from epistemically unreasonable ignorance of the wrong-making features of their act, even when the act does not express ill will.

In response, I provide a theory of culpable ignorance, and blameworthiness more generally, on which ill will is not necessary for blameworthiness. My theory, which I will call Rational Capacitarianism, entails that when an agent acts wrongly from factual ignorance, she is blameworthy if and only if her ignorance is caused by a failure to exercise a rational capacity—a capacity to recognize reasons for believing and desiring and to be generally responsive in one's beliefs and desires to one's assessment of reasons. Failures of rational capacities are relevant because only such failures can be attributed to the agent in a distinctive way.

The theory of blameworthiness contrasts with both quality of will theories of blameworthiness and Capacitarian theories which ground blameworthiness in the failure of capacities but not rational capacities specifically.

DECIDING FOR OTHERS:
An Expressivist Theory of Normative Judgment

(Forthcoming in *Philosophy and Phenomenological Research*)

ABSTRACT

This paper develops a new form of metaethical expressivism according to which the normative judgment that X should Φ consists in a decision that X Φ. When the judgment is first-personal—e.g., my judgment that *I* should Φ—the view is similar to Gibbard's plan expressivism, though the state I call "decision" differs somewhat from a Gibbard-style plan. The deep difference between the views shows in the account of third-personal judgments. Gibbard construes the judgment that Mary should Φ as a *de se* plan on the thinker's part to Φ if she turns out to be Mary (the Subtle View). I construe the judgment as a decision for Mary *that Mary φ* (the Simple View). The main argument for Simple Plan Expressivism is that it solves problems for Gibbard's approach, resonates with a new and interesting moral psychology, and better makes sense of certain independently plausible constraints on normative judgment. In the end I argue that this account of normative judgment has implications for first-order ethics, implying in particular that rational egoism as standardly formulated is incoherent.

1. Introduction

In their effort to provide a systematic account of normative thought and discourse, expressivists have for the most part taken first-personal normative judgments as their primary analysanda. Suppose after my husband cheats on me, I conclude: "I should leave my marriage." According to expressivists, my first-personal judgment that I should leave my marriage consists in a desire-like attitude toward leaving my marriage: e.g., a preference (Dreier 2006, Silk 2015[1]), an intention, or a *sui generis* conative state. For instance, on Allan Gibbard's influential view (2003), my judgment consists in a plan or decision to leave my marriage, a state that settles for me the practical matter of what to do.

My focus in this paper is not moral judgments, but judgments about what people "should" or "ought" to do *all things considered*—the sort of normative judgments that close deliberation and which we express when we give general normative advice. Expressivists aim to capture the practical, motivating character of normative judgment, and when it comes to first-personal judgments about what we ought to do, an expressivist analysis is attractive, since these judgments have an obvious action-guiding role. However, we also make third-personal normative judgments: judgments about what other people should do, all things considered. Gossiping with you in a coffee shop, I might say, "Mary should leave her marriage." This judgment is about what Mary should do, rather than what I should do. Since expressivists must account for the full range of normative judgments, they need an account of these third-personal judgments, and that turns out to be surprisingly difficult to provide.

The simplest view the expressivist could take is that just as my first-personal judgment that I should leave is a desire-like attitude toward leaving my marriage, my third-personal judgment that Mary should leave is a desire-like attitude of the same type toward *Mary*'s leaving her marriage. For instance, if my judgment that I should leave is a preference for my leaving over not leaving, my judgment that Mary should leave is a preference for Mary's

[1] Silk provides a sophisticated preference-based expressivist account of deontic vocabulary generally which may be applied to the sort of all-things-considered "should" judgments I am concerned with.

leaving over her not leaving. If my judgment that I should leave is a decision for me (a decision for me to leave), my judgment that Mary should leave is a decision for Mary: a decision that Mary leave; and so on. Call this the

> **Simple View of Third-Personal Judgments:** Just as my first-personal judgment that I should Φ is a desire-like attitude toward my Φing, my third-personal judgment that X should Φ is a desire-like attitude of the same type toward X's Φing.

The Simple View has a virtue: It is uniform, in the sense that it represents first-personal and third-personal judgments as states of exactly the same kind, differing only in the content of the underlying attitude. But the Simple View is widely rejected, in part because it appears to misdescribe the motivational role of third-personal judgments. On the Simple View the third-personal judgment is a plan, or a desire, or a preference on my part concerning Mary's conduct; but any such state would have to motivate the subject to promote its object, or so it seems. And the trouble is that third-personal normative judgments are not in general motivating in this way. If Mary is my friend, the judgment might come along with something like a desire to help her leave. But I can make the judgment even if Mary is a stranger, or an enemy, in which case it will not incline me to help her leave. This problem arises for any version of the Simple View. But for proponents of the main contemporary form of expressivism, plan expressivism, there would be a further problem. The plan expressivist says that normative judgments are decisions. So, if she adopts the Simple View, she will say that my judgment that Mary should leave is a decision on my part *that Mary leave*. But there seems to be no such thing as deciding for another person. So the plan expressivist looks to have two good reasons for rejecting the Simple View.

The main alternative expressivist treatment of third-personal judgments involves identifying them with *conditional* first-personal attitudes (Gibbard 1983, 2003). My judgment that Mary should leave may not motivate me to help her leave. However, it would, plausibly, motivate me to leave my marriage *if* I were in circumstances just like Mary's in all relevant

respects. This suggests the desire-like attitude expressed by a third-personal judgment is conditional and first-personal: it is a desire-like attitude towards my leaving if, hypothetically, I am in Mary's situation. On Gibbard's view, it is a conditional plan: a plan to leave should I find myself in Mary's exact predicament.

Suppose Holmes is reasoning about what Mrs. Hudson should do. Holmes cannot decide *for* Mrs. Hudson, says Gibbard. But he can form a contingency plan for what to do in her situation:

> [I]f the two aren't in communication tonight [Holmes] knows that his own thinking can in no way settle the question of what to do in Mrs. Hudson's own mind. She can't possibly act on his thinking, the way Holmes later on can act on the planning he's now engaged in. Still, this is in a way a kind of hypothetical planning: he thinks what to do if in a hypothetical circumstance. He's thinking as if he could plan what to do if in Mrs. Hudson's plight … (Gibbard 2003: 50).

Even though Holmes knows that he will never be Mrs. Hudson, he can still make a hypothetical plan for himself for what to do in her shoes, and according to Gibbard, this is what his third-personal judgment that Mrs. Hudson should Φ consists in. Call this the

> **Subtle View of Third-Personal Judgments**: My first-personal judgment that I should Φ is desire-like attitude toward my Φing. My third-personal judgment that X should Φ is a desire-like attitude toward *my Φing in a situation exactly like X's*.

The first part of this paper is critical: I argue that Subtle View of third-personal judgments is inadequate. In the second half, I defend an approach to third-personal judgments within a plan expressivist framework according to which third-personal judgments are unconditional plans or decisions. Specifically, I argue for *Simple Plan Expressivism*, according to which the third-personal judgment that X should Φ is a decision for X *that X Φ*—a decision whose focus is someone else's conduct. This account avoids all the objections to the Subtle View that I set out in the first half the paper.

This is a surprising view— the notion of deciding for another person seems unfamiliar. But, I argue, decisions for others are operative in certain widespread social practices like

delegated decision-making, in which one person allows another to settle a practical question for her. Thus, decisions for others are not only possible, they are commonplace. Moreover, I argue that decisions for others do not motivate the psychological subject of the decision (the decider) to promote their object, but rather function well when they motivate the *executive subject* of the decision—the agent whom the decision is a decision for (Roth 2004). The account is thus consistent with the motivational asymmetry noted earlier, according to which first-personal judgments motivate the subject of the judgment to promote their object, whereas third-personal judgments in general do not. Moreover, the view has the interesting and intuitively plausible consequence that the motivational role of third-personal judgments is essentially social, aimed at guiding the actions of others.[2] I will not pretend to offer a full-dress defense of my approach in this paper. My aim is to set out the view, to highlight its attractive features, and to show why it is superior to Gibbard's account of third-personal judgments.

Finally, I argue that this account of normative judgment has implications for first-order ethics. If my third-personal judgments are decisions I make on behalf of others, then they must harmonize with one another and with my decisions for myself in the same way in which my decisions about what to do now must harmonize with my decisions about what to do at other times. This coherence constraint on decisions restricts the normative judgments we can coherently make, implying in particular that rational egoism—the view that everyone should benefit himself—is not just false but (in a sense I specify) inconsistent. The view thus promises a route from metaethics to ethics and holds out the prospect that reflection on the nature of normative judgment may yield substantive constraints on what we should do.

2. Problems for the Subtle View

[2] Of course, there are a variety of rich and highly developed descriptivist, truth-conditional semantics for "should" which support cognitivism about these judgments. See, for instance, Wedgwood (2006, 2007: ch. 5, 2016), Finlay (2009, 2010, 2014), and Worsnip (2019). Unfortunately, I shall not be able to address any of these cognitivist alternatives here. My aim in this paper is to investigate the best response to the problem of third-personal judgments within an expressivist framework.

I will illustrate the problems with the Subtle View against the background of Gibbard's plan expressivist theory since Gibbard provides the most developed version of the Subtle View. However, the objections apply to any expressivist view which identifies third-personal judgments with conditional first-personal attitudes, like conditional desires or preferences.

Before setting out my three main objections to the Subtle View, I will state a general concern. As noted, Gibbard holds that normative judgments are plans, and in particular that my third-personal judgment that Mary should leave is a plan for me to leave if I find myself in Mary's situation. Many have wondered about the psychological plausibility of the view that third-personal judgments are plans or preferences for remote scenarios (e.g., Cullity 2007, Gregory 2017). Why would we bother forming plans for situations that we know will never arise? According to Gibbard, we engage in such planning as a kind of "rehearsal for life" (Gibbard 2003: 52)—to settle in advance what we will do in various contingencies. The capacity for third-personal judgment emerged in human beings because this sort of rehearsal proved advantageous.

However, the relevant scenarios are often extremely remote; we can make normative judgments about what the president of Venezuela ought to do, or what a character in a fantasy novel ought to do. It is hard to believe that our hominid ancestors derived a practical benefit from by simulating decision-making in situations they will certainly never face. Especially noteworthy is the fact that on Gibbard's view, Holmes is not merely planning to Φ in a situation qualitatively identical to Mrs. Hudson's, but planning to Φ if he is *Mrs. Hudson* in Mrs. Hudson's shoes (Gibbard 2003: 51). If he judges she should leave, he is issuing himself the (conditional) directive: "if you're ever Mrs. Hudson in the exact situation she's in right now, leave." Since it is impossible for Holmes to be Mrs. Hudson, this plan has no chance of being relevant to a situation Holmes finds himself in—not even a qualitatively similar situation (since Holmes would still be Holmes in this situation).

While I share these authors' skepticism on this score to a degree, I think this objection is not decisive. Perhaps we have come to find such contingency planning "fascinating in itself, entirely apart from the ways the practice might later come in handy" (Gibbard 2003: 53). But

psychological plausibility is not the only concern for the Subtle View; so let us turn to the other objections.

2.1. The Problem of Interpersonal Disagreement

My first objection is that the Subtle View does not give a satisfactory account of interpersonal normative disagreement. Suppose you and I are gossiping together about Mary in the coffee shop and I judge that Mary should leave her marriage, while you judge that it is not the case that she should leave. Our normative judgments obviously disagree, so the mental states that correspond to them must disagree. Moreover, our normative judgments disagree in the strong sense that they are inconsistent with one another. This is a datum that any account of normative judgment must capture.

On Gibbard's view, my judgment that Mary should leave is a plan that requires *me* to leave in Mary's situation, whereas your judgment that it's not the case that Mary should leave is a plan that permits *you* not to leave in that situation. Now if a single person made both these plans, his plans would be inconsistent, in the sense that one would forbid an action that the other permits. At least, this is what plan expressivists must say in order to account for why it would be inconsistent for a single agent to judge both that Mary should leave and that it's not the case that Mary should leave. But a special problem arises in the case of interpersonal disagreement. When I make the first plan and you make the second, my plan requires *me* to leave in Mary's shoes, whereas your plan says nothing about my conduct, and so neither forbids nor permits nor requires *my* leaving in Mary's shoes. The disagreement thus goes missing; my plan does not forbid any course of action that your plan permits.

Of course there is a sense in which our plans "disagree" on the Subtle View. We make *de se* plans that would be inconsistent if a single person made them. But that's analogous to the sense in which we "disagree" if I judge that I'm happy and you judge that you are not. These judgments "disagree" in the sense that they respectively self-ascribe properties that no single person could have. But of course they don't really disagree. If I say "I'm happy" and you say "I'm not", though we are accepting and rejecting (respectively) the same centered

proposition, we do not contradict one another in any sense. Gibbard's view entails that when I say "Mary should leave" and you say "No, she shouldn't", we disagree only in this bogus sense. Conflicts in *de se* contents do not amount to disagreement.[3]

The worry that I am raising is not the familiar "negation problem" for expressivism (Unwin 2001, Schroeder 2008). The negation problem is the task of explaining why it is inconsistent even for a single individual to judge both that X should Φ and that it's not the case that X should Φ. Gibbard has a solution to the negation problem. As both Gibbard and I understand them, plans are states that reject or permit actions. A plan requires an action if it permits only actions consistent with it; a plan forbids an action if it permits only actions inconsistent with it. When an agent judges that X should Φ, she accepts a plan that requires her to Φ in X's shoes; when she judges that it's not the case that X should Φ, she accepts a plan that permits her not to Φ in X's shoes. And it can be shown, given plausible definitions, that these plans are inconsistent.[4]

The problem I am pressing is that when you and I disagree about whether Mary should leave, our Gibbard-style plans are not inconsistent in the same way: Yours requires *you* to leave in Mary's shoes; mine permits *me* not to leave. These are *different* plans, but this sort of difference does not amount to disagreement. If I plan to go to the beach today and you plan

[3] See Dreier (2009) for a further discussion of expressivists' problems with disagreement, including a discussion of Gibbard. For further discussion of why, in general, conflicts of de se contents do not suffice for disagreement, see Beddor (2018); Egan (2010, 2014); Richard (2011), and Ross and Schroeder (2013). See Ridge (2013) and Beddor (2018) for accounts of disagreement which might be marshalled in defense of Gibbard.

[4] A Gibbard-style plan is a set of hyperplans, where a hyperplan specifies, for each centered world, one or more actions permitted to the center. A plan permits X to Φ if every element of it permits the center to Φ if the center is X; a plan requires X to Φ iff it does not permit the center not to Φ if the center is X. Plans are inconsistent when their intersection is empty. Since there is no hyperplan that both requires the center to Φ if the center is X and permits the center not to Φ if the center is X, the plans associated with "X should Φ" and "It's not the case that X should Φ" are inconsistent and any single person who accepts both is thereby inconsistent. But note, since these plans are de se contents, built from centered worlds, it will not follow from the fact that you accept one plan, and I accept an inconsistent plan, that we disagree; conflicts in de se contents do not amount to disagreement. That is why the disagreement problem is different from the negation problem. As many have pointed out, Gibbard's formal solution to the negation problem must be supplemented with an independent account of what it is for one state of mind to disagree with another (Schroeder 2008, Dreier 2009). My point is that even if we grant that such an account can be provided, it would not solve the *interpersonal* disagreement problem.

either to go to the beach or stay home—embracing a plan that permits either—we have different plans, but there is nothing we disagree about since your plan concerns how you are to spend the day, and mine says nothing about that.

2.2. The Wrong Subject Matter Objection

My second objection to the Subtle View is that it assigns the wrong subject matter to third-personal normative judgments.

When I'm gossiping with you about whether Mary should leave her marriage, we are talking about Mary and only Mary. My judgment that Mary should leave is focused on Mary in the same sense in which my ordinary descriptive belief that Mary is happy has Mary as its sole focus. The judgment is not at all about me, just as my belief that Mary is happy is not at all about me.

The Subtle View entails that my third-personal judgment is a *de se* attitude, and is therefore partly about myself. I say "partly," because the attitude does make reference to Mary. As Ridge points out,

> [I]ndexical attitudes can take other agents within their scope in the following way: I can prefer that if I were in the other agent's circumstances that certain states of affairs not obtain. The attitude is still indexical in that it is cast in terms of what I want for myself if I were in other person's circumstances. At the same time, it makes a kind of sense of the idea that the judgment is nonetheless in another sense about the other agent in question (Ridge 2011: 48)

However, the Subtle View cannot vindicate the idea that third-personal judgments are solely about the agent and not the subject of the judgment. Consider the difference between my belief that Zoe is nervous and my belief that if I were Zoe, I would be nervous. While the former is solely about Zoe, the latter is partly about me. The Subtle View thus makes judgments about others partly self-regarding in just this way, and that is false to the phenomenology of normative judgment.

One might reply that actually Gibbard gets it right—that third-personal normative judgments are in fact about oneself—by noting that when we give advice, we often say things like "if I were you, I wouldn't do that." (Indeed, Gibbard relies heavily on this point). However,

this is weak evidence for the claim that the judgment expressed by "you should Φ" is equivalent to the judgment expressed by "if I were you, I would Φ". The fact that I think about what I would do if I were you in giving advice is easily explained in other ways. For instance, it is plausibly explained by the fact that it is often easiest to figure out what someone should do in a situation by thinking about what we would do. This is true of factual questions as well. If I want to figure out whether Zoe is nervous, I might do so by thinking about whether I would be nervous in Zoe's situation. In a context in which I'm asked whether I think Zoe is nervous, I might say, "Well, if I were Zoe, I'd be nervous." But this is hardly evidence for the claim that the judgment that Zoe is nervous is identical to the judgment that if I were Zoe, I would be nervous.

2.3. The Conflation Objection

The third objection is that the Subtle View conflates normative judgments that are distinct, in the sense that it is possible to hold one without holding the other (Gregory 2017[5]). Consider the following two judgments: My judgment that Mary should Φ and my judgment that I should Φ if I am Mary in Mary's circumstances. The Subtle View presumably gives these judgments the same analysis: they are both desire-like attitudes toward Φing if one is Mary in Mary's circumstances. In other words, according to the Subtle View, third-personal normative judgments are first-personal normative judgments in disguise. But these judgments are distinct: I can accept one without accepting the other. Gregory does not give an example of a scenario in which someone accepts one but not the other, but we can construct one. Suppose I'm a selfish egomaniac who thinks the following: "In every possible situation, everyone should do what is best for me." I then accept the judgment that if I am Mary in Mary's circumstances,

[5] Gregory presents a version of this objection (Gregory 2017: 43-44). Gregory's version uses the subjunctive conditional, rather than the indicative conditional. I am using the indicative conditional because it is clearer that indicative conditionals are analyzed as contingency plans in a Gibbardian framework. Indeed, Gibbard often states contingency plans in indicative form; for instance, he says that Brutus, in planning to stay home in Caesar's shoes, thinks "If I am Caesar in Caesar's shoes this morning, let me stay home" (Gibbard 2003: 6).

I should do what is best for Mary.[6] However, I reject the unconditional judgment that Mary should do what is best for herself, since I know that I am not Mary. Since I'm an egomaniac, I think Mary should be doing what's best for *me*.[7]

A central desideratum for any non-cognitivist theory is that it avoids conflating distinct normative judgments. If P and Q are normative claims that can be accepted and rejected independently of one another, then the theory should assign P and Q acceptance states which can be held independently of one another.[8] The Subtle View of third-personal judgments flouts this desideratum.[9]

3. The Simple View Solves the Problems for the Subtle View

The Simple View, in contrast, is not subject to these objections. This is because the Simple View does not represent third-personal judgments as implicitly *de se* but rather identifies them with attitudes held directly toward others' actions.

Begin with the *wrong subject matter objection*. Recall that according to the Simple View, my first-personal judgment that I should Φ is

---

[6] The example may be more compelling if we imagine that the egomaniac does not know whether she is Mary. The indicative is infelicitous when the subject knows the antecedent is false. If the egomaniac does not know whether she is Mary, then she accepts the judgment expressed by "if I am Mary in Mary's circumstances, I should do what is best for me" but does not accept the judgment that Mary ought to do what is best for herself, since she does not know whether she is Mary. For more discussion of the conflation point and the egomaniac, see (Ayars & Rosen, forthcoming).

[7] Some philosophers think Qualitative Supervenience—the claim that there are no two possible situations alike in all non-normative qualitative respects but which differ in some normative respect—is a conceptual truth. If they're right, accepting one judgment but not the other is a conceptual confusion. But what matters is whether accepting one judgment and not the other, is *possible*, not whether it is coherent; see Gregory (2017) for discussion. In addition, Gibbard denies Qualitative Supervenience, since he believes one can plan differently for being (say) Jane and for being Mary (Gibbard 2003: 50).

[8] The acceptance state for a statement S is the mental state in which the acceptance of S consists, i.e., roughly, the state that would normally be expressed by a sincere assertoric utterance of S (Dreier 1996)

[9] I concede that this test of the distinctness of the judgments (involving the egomaniac) is not decisive. It certainly seems possible to judge that Hesperus is rising without judging that Phosphorus is rising; but sophisticated solutions to Frege's puzzle sometimes say otherwise (Soames 1989). Still it counts against a theory if it identifies judgments that seem on informed reflection to be modally separable.

(D1) a desire-like attitude toward my Φing

whereas the third-personal judgment that Mary should Φ is

(D2) a desire-like attitude toward Mary's Φing

(in contrast to a desire-like attitude toward Φing in Mary's shoes, as the Subtle View posits). This is a unified view according to which the attitude expressed by "X should Φ" is about X and only X in both the first-personal and third-personal case. Hence, the wrong subject-matter objection is avoided.

The *problem of interpersonal disagreement* is also solved. Someone who thinks Mary should leave disagrees with someone who thinks that it is not the case that Mary should leave in virtue of their having inconsistent desire-like attitudes toward *Mary's* leaving. If the relevant attitudes are plans, the judgment that Mary should leave is a plan *that Mary leave* (i.e., that forbids Mary's not leaving). Someone who thinks Mary should leave has a plan that forbids Mary's not leaving, and someone who thinks that it's not the case that Mary should leave has a plan that permits Mary's not leaving. The plans thus disagree in a straightforward sense. On the Simple View, disagreement between individuals is like disagreement within a single individual. Insofar as expressivists can give a satisfactory account of intrapersonal normative disagreement (see n. 4.), interpersonal disagreement comes for free. This is a major advantage of the Simple View.

Finally, the view solves the *conflation objection*: it distinguishes the judgment that Mary should Φ from the judgment expressed by "I should Φ if I am Mary in Mary's situation," identifying the former with a desire-like attitude toward Mary Φing (e.g., a plan that Mary Φ) and the latter with a desire-like attitude toward Φing if one is Mary in Mary's shoes (e.g., a plan to Φ if one is Mary in Mary's situation). The egomaniac has a desire-like attitude toward benefiting Mary if he is Mary, without having an unconditional desire-like attitude toward

Mary's benefiting Mary. The view thus assigns distinct acceptance-states to these distinct normative judgments.

The Simple View immediately solves the objections to the Subtle View. So it would be nice if the Simple View were tenable. But is it?

4. Simple Plan Expressivism

Whether the Simple View is tenable, I believe, depends on *which* desire-like attitude the expressivist takes normative judgments to consist in. Any view that identifies third-personal judgments with preferences or desires that others do things is a nonstarter; it is just obvious that you can judge that someone should do something without preferring or desiring that she do it. Consequently, in the remainder of the paper, I will defend a view of third-personal normative judgments within a plan expressivist framework according to which third-personal judgments are unconditional *decisions* for others that others do things, as opposed to conditional decisions for the subject to do things in remote situations, as the Subtle View posits. Call this

> **Simple Plan Expressivism:** Just as my first-personal judgment that I should Φ is a decision that I Φ, my third-personal judgment that X should Φ is a decision that X Φ.

Simple Plan Expressivism gives third-personal judgments a uniform analysis, since a first-personal decision to Φ is a special case of a decision-*that*—a decision that one Φ.[10] It is hence a version of the Simple View and avoids all the objections to Gibbard's Subtle Plan-Expressivist View of third-personal judgments.

I will say why I think Simple Plan Expressivism is promising shortly. But first let me make a preliminary clarification. It is important for the plan expressivist to insist on a

[10] English grammar omits the subject in infinitive clauses, so we speak simply of my decision to Φ. But it is clear that every such planning state consists in a system of decisions for the planner to do this or that in such and such conditions.

distinction between the state we are calling "decision" or "plan" and an ordinary intention. (This involves a departure from ordinary language, where to decide to Φ just is to form the intention to Φ). As many critics of Gibbard have noted, some intentions are obviously not normative judgments. We sometimes *pick* between options which we believe to be equidesirable. Facing a rack of practically identical black dresses that I might wear to the cocktail party, I pick the one directly in front of me. But there is no judgment that I *should* pick that one over the others on the rack, since I believe them all to be equidesirable. Call this a *Buridan case*. Sometimes we form intentions akratically, as when I decide to get drunk at the cocktail party while knowing I should not. Assuming I am not of "divided mind,"[11] the akratic intention is not a normative judgment. Bratman calls this the *restriction problem* for plan expressivism: "Gibbard's theory needs a way of restricting which decisions or contingency plans are normatively expressible. It needs to explain why specific decisions (plans) [like the plans in Buridan cases] are not expressible normatively, even though the state of mind with which his expressivism starts is that of decisions (plans) to act in certain circumstances" (Bratman 2006: 708).

In light of the restriction problem, the plan expressivist should not identify normative judgments with intentions at all. Rather, she should identify them with a *sui generis* decisional state whose characterization must proceed, I believe, in functionalist terms. *A decision is the state that concludes practical deliberation*. The essential role of decision (and hence normative judgment) in our psychological economy is not to represent an independent normative fact, as the cognitivist would have it, but to settle or constrain our intentions up to arbitrary picking in a fashion responsive to our judgments about our reasons.[12] We might think

[11] Gibbard argues that akrasia does involve divided mind. If you judge that you ought to Φ but fail to intend to Φ, "there's a part of you that doesn't really think you ought to. You are of more than one mind on whether you [ought to Φ]" (2003: 153). The solution I am offering departs from Gibbard's treatment of akrasia. My solution vindicates the possibility of weakness of will as traditionally understood: intentional action that is contrary to one's (wholehearted) judgment about what one ought to do.

[12] What are judgments about reasons in a plan expressivist framework? Gibbard understands the notion of a reason in terms of the psychological notion of "weighing a factor in favor" of an act. To judge that R is a reason to do A is to "form contingency plans on a certain pattern," a pattern which involves weighing R in favor of doing A in deliberation about what

of a decision as a state that is posterior to deliberation and just prior to intention: a decision closes practical deliberation and then is *executed* by forming an ordinary intention. Sometimes we give ourselves a menu of options to pick from, as when, at the clothing store, I think: "any one of these dresses would be okay." The critical point is that the well-functioning of normative judgments consists (at least in part) in their issuing in an intention in line with the selection, and hence their direction of fit is (at least in part) world-to-mind. This is the sense in which normative judgment is a practical state deserving of the name "decision".[13,14]

Accordingly, Simple Plan Expressivism is not the view that third-personal judgments consist in *intentions* that others do things. The view is that third-personal judgments consist in a *sui generis* state with a certain functional role, which—for reasons that will become apparent—deserves to be called a decision for another.

To know whether Simple Plan Expressivism is tenable, we need to address the objections stated earlier. The first objection was that the very idea of deciding for another person makes no sense. Decisions answer the practical question of what *to* do, and function well when they issue in an intention and thus motivate the agent to act. In the first-personal case, my decision ratifies one or more courses of action. When all goes well, I form an intention by picking one of them. The worry is that nothing comparable is possible in the third-personal

---

to do (2003: 190). See Bratman (2000: 257-258) for view according to which judgments about reasons consist in deliberative policies.

[13] See, e.g., Dreier (2015), who says that "it is part of the essential role of a normative judgment that it motivates" (164). Normative judgments are *supposed* to motivate in some sense (2015: 164); this is what distinguishes their role from that of cognitive states.

[14] Here is an analogy. Suppose two people are on a road trip in an unfamiliar place. The passenger, map in hand, is in charge of navigation. Her task is to instruct the driver where to go. She surveys the reasons for and against taking various routes and deploys these judgments to arrive at instructions to the driver: "turn left here," "take the next exit," "going either right or left would be okay," etc. The driver's role is to execute the passenger's instructions, by setting the car in a certain direction. When all goes well, the driver dutifully follows the passenger's instructions. There is a reasonable sense in which the passenger is deciding where the car goes, even though she doesn't have control of the wheel. The passenger settles the *question* of what route to take, in light of the reasons, by ratifying one or more options; her instructions "function well" when they are appropriately executed by the driver. In this scenario, the passenger is playing the role of normative judgment, and the driver is playing the role of intention.

case. How could I be in a state whose role is to resolve a practical question for *someone else*? My mental states cannot reach into the brain of someone else and cause an intention to materialize. It seems that there is an *own-action condition* on decisions: one can only decide one's own actions. Call this the

> *First Objection to Simple Plan Expressivism:* Other people's actions are not under our control; our mental states do not issue in intentions on the part of others. But a decision is a state whose role is to issue in an intention. So, there is no such thing deciding for another.

The second objection was that Simple Plan Expressivism appears to misdescribe the motivational role of third-personal judgments:

> *Second Objection to Simple Plan Expressivism*: We can judge that someone should do something without being motivated to promote their doing it. But Simple Plan Expressivism arguably implies that anyone who judges that X should Φ is motivated to promote X's Φing, since decisions seem to be states which motivate the subject to promote their object.

In the next section, I will give a more thorough exposition of Simple Plan Expressivism. Then I will show, contra the First Objection, that we can and do settle practical questions for other people, and that contra the Second Objection, Simple Plan Expressivism does not entail that someone who judges that X should Φ is thereby motivated to promote X's Φing.

5. The Objections to Simple Plan Expressivism Reconsidered

Simple Plan Expressivism says that third-personal normative judgments consist in decisions for others, rather than contingency plans to do things in others' shoes. But can we really make decisions for others? And even if we can, why think third-personal normative

judgments consist in such decisions? I will begin by discussing an advisory social practice, which I call *delegated decision-making*, in which decisions for others are operative, and in which these decisions can be expressed using normative terms.

Suppose you and I are at a restaurant, and you know more about the menu and what I would like than I do. So, I ask you to decide what I will order: "Tell me what to get," I say. You deliberate, taking my tastes, my budget and the rest into account; and then voice your conclusion with an imperative— "Get the burrito"—or a normative assertion: "You should get the burrito". Either way, the judgment aims to settle for *me* the practical question of what to order in a fashion sensitive to my choice situation. I asked you to decide on my behalf what I am to get, and you did. If the exchange goes smoothly, I act on your decision. Your decision issues in an intention on my part to get the burrito, much as my own decisions (normative judgments) issue in intentions for me when all goes smoothly.

Of course, the process by which the decision issues in the intention involves additional steps in the third-personal case. You must communicate your decision to me, for instance. There is no magic by which your decisions automatically produce intentions in me. Still, the *role* played by your decisions is in relevant respects the same as the role that my own first-personal decisions play in my psyche: to guide and give rise to action on my part. In delegated decision-making, instead of deliberating myself about what to do, you deliberate for me: thus it is *you* who takes my reasons into account and settles the practical question of what I am to do by providing the menu. So, it is your mental state which fulfills the relevant role.

The essential feature of delegated decision-making, as I have characterized it, is that person A delegates a decision to person B who then *decides* (on behalf of A) *that* A is to do something. B hence makes a real decision, which has as its fulfillment condition A's performing some action. Not all instances of reason-giving or persuasion involve decisions for others in this sense. If you merely point me to a consideration in favor of getting the burrito, you have provided me with a reason (by bringing it to my attention), but you have made no decision. You are not in a state whose aim is to *settle* my intentions. You are only offering input.

There are two main sources of support for the claim that delegated decision-making involves deciding for others in this sense. The first is the role played by the output of delegated decision-making, as just discussed. Your decision functions well when it motivates *me* to act, and hence is plausibly a decision for me.

Another piece of evidence that you are making real decisions that refer to my actions in this context. Notice that certain *rational norms* characteristic of decisions appear to govern your states.[15] Suppose the restaurant only has one burrito left, and that you know this. If you decide on my behalf that I am to get it, you cannot consistently also decide to get the last burrito for yourself. If after telling me I should get the last burrito you say to the waiter, "I'll get that last burrito," I would be puzzled, and could reasonably complain: "I thought you said *I* should get the burrito." This makes sense if your state is a real decision for me on your part, since if it is a real decision, it must harmonize with your other decisions, including your decisions for yourself, in the same way that your decisions about what to do now constrain your decisions about what to do at other times. If you decide to Φ, and also to Ψ, it must be possible for you to both Φ and Ψ. And more generally, if you decide (in our technical sense) that X is to Φ and that Y is to Ψ, then it must be possible for X to Φ and Y to Ψ. (This *co-realizability constraint* is the aspect of this view that makes trouble for rational egoism, as I will show later).

Notice that on the Subtle View of what your normative assertion ("you should get the burrito") expresses, there is no incoherence in making the assertion and then deciding to order the burrito yourself. Your judgment, on the Subtle View, is a conditional decision to order the burrito in my shoes. Since you are not in my shoes, you are rationally free to form the conditional plan to get it in that (unrealized) situation and then to decide to do something

[15] Roth argues on similar grounds that we can form intentions that others do things (Roth 2004, 2014, 2016). I am, however, making a claim about the state associated with normative judgment, which is not intention in the ordinary sense.

incompatible in the situation you actually face. This counts in favor of the present view, against the Subtle View.[16]

Contra the First Objection—the complaint that there is no such thing as settling a practical question for someone else—the example of delegated decision-making shows that we can and do make bona fide decisions that refer to others' actions, and these decisions are normatively expressible. We can intelligibly address, and sometimes resolve, the practical question of what *someone else* is to do. Some of the considerations I am invoking in support of this should be familiar. Philosophers have argued, on similar grounds, that we can form *intentions* which refer to other peoples' actions (Searle 1990; Gilbert 1990, 1996; Bratman 1992, 2013) and that the rational norms characteristic of intentions can extend across people (Roth 2004, 2014, 2016, 2018). I am, however, advancing a novel claim about the state associated with *normative judgment*. As noted earlier, for the plan expressivist normative judgment is not intention in the ordinary sense. My claim is that we can answer practical questions for others, and thus "decide for others" in the heavyweight sense of *decision* whose mark is aptness for expression in normative terms: e.g., "you should get the burrito".

Delegated decision-making is rare, and if this were the only clear context in which we can make decisions for others which fulfill their role, the point might be of limited interest. But this is not the only such context. Plausibly, the paradigmatic normative exchange is *co-deliberative*, rather than purely deferential as in delegated-decision making. Normative

[16] To convey the feel of how the view is different from Gibbard's, consider an example in which someone decides that she will Φ if she is in X's situation without deciding for X that X Φ. You could decide to get the burrito in my shoes while withholding judgment on what I should do. You might express this by saying, "I don't know what you should do—I'm not going to decide for you—though if *I* were you, I'd get the burrito." Your conditional decision would commit you to a motivation to get the burrito in my shoes. But it would not commit you to deciding not to interfere since it falls short of an endorsement of my getting the burrito (assuming you genuinely withhold judgment on what I should do).

However, someone who judges that if she were in X's shoes, she should Φ but refuses judge that X should Φ strikes as cagey or unreasonable. This is because normative judgments are not supposed to be tied to particular people. Any reason to favor your getting the burrito in my exact situation (which includes all of my qualitative properties) would also be a reason to favor my getting the burrito. Judging that you should get the burrito in my shoes *commits* you to judging that I should get the burrito, and this is why the assertion described sounds odd.

judgment paradigmatically involves reasoning *together* with someone about what she is to do. That is, I concur with Manne's (2014) suggestion that normative judgments are most at home in the "interpersonal" mode of interaction, i.e., within a certain kind of commonplace social practice in which one reasons *with* someone about what she is to do. If I judge that you should Φ and tell you what I think and why I think it, you may come to agree with my judgment. In agreeing with my judgment, you make a decision for yourself: a decision to Φ.[17] If all goes well, you then act on the (joint) decision. My decision for you fulfills its role indirectly by securing agreement with you.

Of course, we can make third-personal normative judgments outside these "cooperative" social practices. We can make normative judgments about strangers and enemies, with no expectation that they will be acted on. We can make normative judgments privately, without expressing them at all. Simple Plan Expressivism says that these normative judgments consist in the *same underlying state* as is present in delegated decision-making and co-deliberation—a decision for another—even if there has been no delegation of authority and even if there is no cooperative activity. But is it plausible that normative judgments consist (across the board) in decisions for others?

There is, I believe, an excellent preliminary case for this. As noted, we can express decisions for others in the context of delegated decision-making (and co-deliberation) using "should" statements ("you should get the burrito"). But all all-things-considered "should" statements have the same meaning, and so express the same state, regardless of the context in which they are uttered or accepted. That suggests that even in these "non-cooperative" contexts, the normative judgments these statements express are best understood as decisions. The additional intentions present in delegated decision-making and co-deliberation—e.g., the intention to communicate the conclusion of the relevant reasoning—are not constitutive of the *decision*. The decision is simply the *practical conclusion* which can be expressed in normative

[17] Note that if you concur with my judgment by making a decision for yourself, the state you agreed with must also have been a decision for you.

terms, and this presumably the same state we are in when making a private judgment about what someone should do, or overt judgments in contexts where the judgment stands no chance of being acted on.

The proposal, however, raises a worry. The view assumes that we can coherently decide (in our technical sense) on something over which we know we have no control. Suppose I make a normative judgment about someone whose conduct I cannot affect. I judge that my husband should stop cheating on his partners, knowing he will not. Simple Plan Expressivism says that the judgment is a decision for my husband whose constitutive "point" is to affect his intentions. Since my normative *judgment* may be perfectly coherent (and apt), the plan expressivist must suppose that such knowably inefficacious *decisions* can be coherent (and apt). But is this plausible?

The first thing to note is that every plan expressivist must grapple with this problem of explaining the coherence of such decisions. As Gibbard observes, *first-personal* normative judgments—which the expressivist identifies with *de se* plans—are also often (knowably) inefficacious, though they may be exactly what the situation calls for. A binge alcoholic, knowing he will be driven by compulsion to drink on Saturday, can still judge—appropriately—that he should not. Gibbard claims that asked Wednesday for a decision whether to get drunk in his situation to come, he can decide not to, though "he can't make this hypothetical decision effective" (50). Consider also judgments about the past. My judgment that I should not have had that third drink last night can be appropriate even though the hypothetical plan that constitutes it has no chance of affecting my past conduct, as Gibbard acknowledges (50-51).

Since every plan expressivist must concede that some decisions (normative judgments) are inevitably inefficacious in this way, the fact that my view saddles us with many such plans is not a reason to favor Gibbard's Subtle View of third-personal judgments over Simple Plan Expressivism, which is the question I am concerned with in this article.[18] No doubt, the plan

[18] One might, of course, think that the plan expressivism's commitment to the aptness of such decisions makes plan expressivism untenable, but my aim in this paper is to find the most promising version of plan expressivism, not to defend the expressivist against the cognitivists.

expressivist will emphasize that normative judgments are not intentions, but a *sui generis* state, which is subject to a distinct set of norms. For instance, while intention is subject to a Settling Condition (it is irrational to intend something you do not believe you can settle),[19] the norm on decision appears to be weaker: decisions, in our technical sense, should be such that nothing debars them from playing their designated output role *if circumstances arise where they are called upon to do so*. My husband *could* act on my judgment/decision (it is in his power to stop cheating on his partners), so it satisfies this weaker condition.[20]

Still, one might be skeptical: how could a state be apt if has no prospect—in a particular case—of fulfilling its normal output role? In response to this general worry, the plan expressivist, I believe, should lean on the observation that it is true of many types of states that they can be fully apt or warranted in their own terms—however these notions are understood—even when they cannot possibly play their official role.

Consider belief. In one sense, the aim of belief is to represent things as they are, so any true belief serves the purpose. But in another sense, the aim of belief is to guide action by channeling intentions from ends to means in ways that benefit the agent. That is arguably the purpose for which beliefs as a class exist, so a belief that represents the world accurately but fails to affect how the agent behaves has not played this type-defining role. But while the first sort of role determines a correctness condition for every token state, the second does not, though it may yield norms of other kinds. Beliefs are not inapt just because they stand no chance of guiding action. The second is closer to the conception of the role of a state I have been invoking: roughly, its "point" in our psychological economy, or what it does for agents in the course of deliberate action. The same, I say, goes for decisions. An inefficacious decision that is otherwise justifiable is like an inefficacious true belief: though it fails to do what states of that kind exist in order to do, it is not defective in itself.

Now let me show why Simple Plan Expressivism is not subject to the Second Objection, concerning motivational role. Simple Plan Expressivism, I submit, does not imply that anyone

[19] Bratman (1997)
[20] This output norm corresponds to an "ought implies can" constraint.

who judges that X should Φ is thereby motivated to promote X's Φing. To see why, consider again the functional role of third-personal judgments/decisions. According to the view sketched above, their role is to influence the intentions of someone else. Following Roth (2004), we can distinguish between the psychological subject of a decision—the individual who issues the decision—from the "executive subject" of the decision, the individual whose "job" it is to execute the decision. The executive subject of your first-personal decisions is you. But the executive subject your decision that I get the burrito is me. A decision generates a motivation only when it leads to an intention; but even when all goes well, a decision only leads to an intention in its executive subject. So while your decision for me that I get the burrito is poised to generate an intention and a motivation in me (and likely will if it is expressed), you might not care at all whether I fulfill your decision, nor are you rationally committed to caring, and so need not be motivated to help me get the burrito. After all the judgment/decision is meant to settle a practical question *for me*, not for you. Its functional role is *social*, extending "outside the head." Since the role of third-personal decisions is to motivate the executive subject of the decision rather than the psychological subject, Simple Plan Expressivism is consistent with the phenomenological datum that my judgment that Mary should leave her marriage does not imply an inclination on my part to help Mary leave.

This is a clear advantage of identifying the third-personal judgment that X should Φ with the decision *that* X Φ as opposed to the decision to do whatever one can to bring it about that X Φs. This nearby view would also avoid the objections to the Subtle View I set out in the first half of the paper, but it would be manifestly inconsistent with the motivational role of these judgments, as Ridge (2018) observes. It also shows why it is critical to identify third personal judgments with decisions that others do things rather than preferences or desires that others do things, since the role of desire and preference is always to motivate the *bearer* of the preference or desire to bring about the wanted state of affairs.

Simple Plan Expressivism, I believe, resonates with a compelling view of the role of normative judgment, according to which the "point" of these judgments—their raison d'être— is to control intentions, both intrasubjectively and intersubjectively, in light of

reasons. As agents, we not only have the ability to step back from our own impulses and ratify or not the course of action they seem to favor, but also to guide the intentions of other agents in our milieux by deliberating from their point of view and ratifying or not courses of action for them. In the intrasubjective case, we often do a better job of acting on our reasons if we step back and reason from them, bringing the full range of our background knowledge to bear. Similarly, since it will often happen that others know more than we do and are less subject to some of the pathologies of decision than we are when we deliberate in our own case, it can be valuable for others to deliberate on our behalf, to express those decisions, and for us at least sometimes to act on them. The hypothesis I am advancing is that normative discourse exists so as to provide the resources for exercising this valuable capacity for intra- and intersubjective regulation of intention by decision for ourselves and others.

Contrast this with Gibbard's view, according to which the usefulness of third-personal judgments is to "rehearse for life" by settling in advance what we will do in remote counterfactual scenarios (Gibbard 2003: 52). Gibbard's view seems to misdescribe the well-functioning of third-personal judgments. It characterizes these judgments as most "at home" in first-personal, individual planning activity, rather than in the *social practices* in which they seem to do most of their distinctive work. To be sure, Simple Plan Expressivism does not explain why we make normative judgments outside the relevant social practices, and I will not attempt to provide such an explanation here. My aim here is to offer an alternative account to Gibbard's of the *fundamental* point or raison d'être of third-personal normative judgment.

To sum up the argument of this section: Simple Plan Expressivism avoids the objections to Gibbard's conditional attitude view discussed in the first half of the paper (the problem of disagreement, the wrong subject matter objection, and the conflation objection), provides a uniform account of first and third-personal judgments, and can be supported by a compelling account of the point of normative judgment which is worth exploring in its own right. The two most pressing objections to the Simple View—the First and Second Objections—can be answered, as I have shown. This is an excellent case for preferring Simple Plan

Expressivism to Gibbard's Subtle View, and for taking it seriously as a better articulation of the plan expressivist's main idea.

6. Rational Constraints on Normative Judgment

In this section, I will show that whether we adopt Simple Plan Expressivism or Gibbard's Subtle View of third-personal judgments has implications beyond esoteric debates in metaethics. In particular, Simple Plan Expressivism entails rational constraints on normative judgment that do not follow from the Subtle View. This makes the choice between the two views consequential for normative ethics. I will suggest that these rational constraints are independently plausible. Hence, they ground an additional argument for Simple Plan Expressivism over the Subtle View.

Recall the burrito example: If you decide on my behalf that I'm to get the last burrito, you cannot coherently decide to get it yourself. Simple Plan Expressivism explains this by identifying judgments with decisions and noting that decisions are subject to a co-realizability constraint: it is irrational to decide that A Φ and B Ψ if it is impossible that [A Φ and B Ψ]. Given this principle, Simple Plan Expressivism entails a range of non-trivial coherence constraints on normative judgment. In particular, Simple Plan Expressivism entails that if you judge that X should Φ, you are committed to not interfering with X's Φing. The judgment that X should Φ is a decision that X Φ; the decision to interfere with X's Φing is a decision to do something incompatible with X's Φing. Since non-co-realizable decisions are incoherent, the corresponding normative judgments are likewise incoherent. So making the first commits you to not making the second.[21]

This is the element of the view that entails that rational egoism as standardly formulated cannot be coherently accepted. Rational egoism, on the standard formulation, says

[21] Simple Plan Expressivism does not entail by this route that you are committed to promoting X's Φing if you judge that X should Φ, since the decision not to promote X's Φing is in general co-realizable with the decision that X Φ. The view thus vindicates Wallace's (2009) idea that the reasons of others provide us with reasons not to interfere with their rational choices without at the same to providing us with reasons to promote their ends.

one should perform some action if and only if and because it maximizes one's self-interest. Suppose you and I are competing for some object. It would maximize my self-interest to acquire it for myself, and it would maximize your self-interest to acquire it for yourself. If I am an egoist, I am committed to the following judgments:

(J1) the judgment that I should take the object for myself (because that would be best for me); and

(J2) the judgment that you should take the object for yourself (because that would be best for you)

Simple Plan Expressivism tells us that I must therefore have made the following decisions:

(D1) a decision for me that I take the object for myself

(D2) a decision for you that you take the object for yourself

But these decisions are incoherent in a straightforward sense. Just as a set of beliefs is coherent (consistent) only if they can all be true together, so a set of decisions is coherent only if they can all be realized together.[22] (D1) and (D2) are incoherent in this sense, so Simple Plan Expressivism entails that (J1) and (J2) are incoherent. And of course egoism is not the only problem case. The view entails the incoherence of any normative view which prescribes incompatible courses of action for individuals. More specifically, Simple Plan Expressivism entails that that the following *interpersonal joint satisfiability* principle must be true: if A should Φ and B should Ψ; then it is possible that [A Φ and B Ψ].[23]

[22] This argument resembles Medlin's (1957) argument against egoism which also takes non-cognitivism as a premise. Medlin argued that the egoist has incoherent preferences in situations like the one described. See Dreier (1996) for discussion.

[23] Most philosophers already accept the following *intrapersonal* joint satisfiability principle: if A should Φ and A should ψ; then it is possible that [A Φ and A ψ].

This amounts to a formal refutation of egoism taking Simple Plan Expressivism as a premise. A possible reaction is to see this as a favorable, and possibly deep, implication of Simple Plan Expressivism. While egoism is now widely regarded as a coherent if mistaken view, given the history of the subject, we should see this consensus for what it is: a counsel of despair, an idea we've gotten used to only because we have not been able to do better. Philosophers throughout history have spilled much ink attempting to refute the egoist by disclosing some deep formal difficulty (Sidgwick 1974, Moore 1903, Kant 1785, Medlin 1957, Baier 1965, Korsgaard 1996). We can all agree, as a substantive matter, that we should be altruistic to some extent. But *why should* one act, at least to some extent, for the sake of others and not only for one's own sake? The strongest possible answer would be a persuasive argument that would show egoist to be untenable or incoherent or otherwise unacceptable *by the egoist's own lights*: the sort of argument that would lead any reflective, truly rational egoist to revise his view in favor of some form of altruism. [24] Simple Plan Expressivism appears to give us such an argument, and if it did, that would be a good thing.

However, this line of reasoning is blunted by the fact that there is a view in the vicinity of rational egoism that is coherent on Simple Plan Expressivism: the view that everyone should *try* to do what is best for himself. It is not incoherent to decide that A try to Φ and that B try to Ψ even if it is impossible that A Φ and B Ψ, since it is always possible for each agent to try. The egoist can simply retreat to this modified view. So there is no quick argument from Simple Plan Expressivism to a positive requirement of altruism.

But even if his is a disappointment, it is also a good thing. The coherence of these "trying" judgments provides a means of answering an objection to Simple Plan Expressivism and the rational constraints on normative judgment that it implies. After all, one might see the rational constraints entailed by the Simple Plan Expressivism not as a favorable implication of it, but as a serious blight. The idea that normative judgments must be jointly satisfiable sounds right when the judgments are expressed and the context is cooperative, as in delegated

[24] Arguments in this vein include those of G.E. Moore (1903), Kant, and Medlin (1957).

decision-making and co-deliberation. But it seems to clash with commonsense in other contexts. A burglar judges that the homeowner should call the cops. Cutting the phone lines, he knows, would preclude her from doing this. Is the burglar then incoherent if he proclaims that he should cut the phone lines? Intuitively not. So while Simple Plan Expressivism neatly explains the requirement of non-interference in delegated decision making, it appears to clash with intuition in other contexts.

But here we can invoke the "trying defense."[25] The trying defense says that what we often mean—or should mean—in contexts in which we say that people should do incompatible things is that each person should *try*.[26] When we say that people should be happy, or cure cancer, or write well, what we ought to say is that they should do what they can to be happy/cure cancer/write well. Analogously, what the burglar ought to say is that the homeowners should try to call the police and that he should try to thwart them, since what is in their power is to try. If judgments like (J1) and (J2) strike us as coherent, this is because we are interpreting them as "trying" judgments.

Indeed, these observations may be turned into an argument for Simple Plan Expressivism. Note that on reflection, even in the absence of theory, there is something odd about (J1) and (J2) when they are not interpreted as "trying" claims. It would be fine to say that I should *try* to take the object for myself and you should try to thwart me. But it sounds odd to say outright that we each *should* take the object for ourselves since that's impossible. Suppose Alice and Bob come to you for advice and ask what they should do if they find themselves in such a situation. It would be odd for you in your capacity as advisor to say: "Well, everyone should do what's best for himself, so you, Alice, should take the object for yourself, and you, Bob, should take it for yourself." That sort of advice sounds not just false but

[25] The trying defense is discussed by Kalin (1970) and Dreier (1996) in response to Medlin's (1957) non-cognitivist argument for the incoherence of egoism.

[26] See Dreier (1996: 415) for a statement of this idea.

incoherent. This suggests that the interpersonal joint satisfiability may well be true. If it is, it is a point in favor of Simple Plan Expressivism that it entails it.[27]

To summarize: The choice between Simple Plan Expressivism and the Gibbard's Subtle View consequential. If Simple Plan Expressivism is true, so is the interpersonal joint satisfiability principle. At first glance, this may seem to be problematic for Simple Plan Expressivism, because judgments like (J1) and (J2) initially appear to be coherent. However, on reflection, there is something odd about (J1) and (J2). Simple Plan Expressivism explains this: (J1) and (J2) are incoherent because they consist in decisions for others which are subject to a co-realizability constraint. This is an additional argument for the Simple Plan Expressivism over both the Subtle View.

7. Conclusion

I have argued that the best expressivist treatment of third-personal judgments of the form "X should Φ" identifies them with a kind of plan or decision *that X Φ*—a decision whose focus is someone else's conduct. This view of normative third-personal judgments—aside from being simple and elegant—avoids the objections to the Subtle View of third-personal judgments set out in the first half of the paper, resonates with a new and interesting functionalist moral psychology, and makes better sense of certain independently plausible constraints on normative judgment.

[27] One might object that the trying defense cannot succeed because sometimes trying is counterproductive. Consider the "paradox of hedonism": we often better succeed at achieving happiness if we do not deliberately seek it. The notion of "trying" can be understood in multiple ways. On one interpretation, to try to Φ is to aim at Φing. But "trying" to Φ can also mean doing what one can to Φ. If trying is interpretated in the latter way, the trying defense can survive the paradox of hedonism worry. To say that everyone should try to maximize their happiness, on this interpretation, is to say that everyone should do what they can to maximize their happiness. This needn't involve aiming at happiness. Thanks to an anonymous reviewer for raising this concern.

AN EXPLANATION FOR THE ESSENTIAL PUBLICITY OF PRACTICAL REASONS

ABSTRACT

In this paper I derive a significant thesis about reasons from an independently motivated non-cognitivist account of normative judgment, in particular, "ought" judgments and reasons judgments. The thesis is that practical reasons are public: If R is a reason for X to do A in C, then R is a reason for anyone not to interfere with X's doing A in C (Wallace 2009). If this form of non-cognitivism is true, then anyone who judges that R is a reason for X to Φ is rationally committed to judging that R is a reason to not interfere with X's Φing. Hence, everyone who makes positive judgments about reasons is committed to every instance of the Publicity Thesis by the nature of normative judgment. If true, the view entails that practical reasons are "essentially" public, though in a sense that differs somewhat from Korsgaard's (1996) or Nagel's (1970).

Note to Readers:

In order to make this a free-standing paper, it was necessary to repeat much of the material from the previous paper and the introduction of this dissertation. The introduction of this paper and the section on the objections to Gibbard — "An (Abbreviated) Case Against the Subtle View" — may be skipped if desired. The most critical new material begins at section IV, "Reasons Judgments and the Publicity Thesis."

1. Introduction

Suppose I am stepping on your broken toe, causing you pain. As Nagel observes in the *The Possibility of Altruism* (1970), you would not merely dislike this; you would resent it. Implicit in your resentment is the supposition that I should stop, and hence, that I have a *reason* to stop. So if your resentment is legitimate, your pain has a dual significance. It gives you a reason to pull your foot away, and it also gives me a reason to (at least) not interfere with your pulling your foot away.

Nagel famously argued that every agent-relative reason implies an agent-neutral one — that every reason for X to Φ is a reason of equal force for everyone to see to it that X Φs. So, for instance, if you have a reason to climb Mt. Kilimanjaro because it is your lifelong goal, this fact is a reason of equal force for me to promote your climbing Kilimanjaro.

This view is extreme, and Nagel has since retracted it.[28] While your reason to climb Kilimanjaro may have some normative significance for me, it is not plausible that I have just the same reason to promote your climbing it as you do. The fact that it is *your* project to climb Kilimanjaro, not mine, makes some difference. Some philosophers now defend a more moderate but still significant claim about reasons (Wallace 2009, Darwall 1983, Korsgaard 1996, Bertea 2017). They contend that reasons are "public": their normative force is always shared. But they deny that a reason for X to Φ is always a reason to promote X's Φing. In Wallace's[29] version:

> **Publicity Thesis:** Necessarily, if R is a reason for X to Φ in C, then R is a reason for anyone to not interfere with X's Φing in C.

[28] Nagel has rejected it on the strength of objections raised by Nicholas Sturgeon in, 'Altruism, Solipsism and the Objectivity of Reasons' (1974: 374–402). Nagel acknowledges this in 'The View from Nowhere' (1986: 159).

[29] As Wallace puts it: "The most basic expression of the normative significance that my reasons have for others is the (pro tanto) requirement of non-interference" (2009: 477).

If you have reason to climb Kilimanjaro, this may not give me reason of equal force to promote your climbing it. (In fact it may give me no reason at all to promote it.) But plausibly, it gives me some reason to not interfere with your climbing it.

The Publicity Thesis is credible in this form. It does not imply that all reasons are agent-neutral, or that we all have reasons to promote the same ends. Even in competitive contexts, others' reasons plausibly generate reasons to not interfere. If you and I are shipwrecked, fighting for the last plank, it may be that each of us should, all things considered, try to get the plank for herself. But even in this context, your reasons still give me reasons. The fact that you need the plank to save your life gives me *some* reason to cede it to you, even if that reason is overridden by the reasons that I have to save myself.

If true, the Publicity Thesis is a deep, important fact about how reasons work. It amounts to the claim that practical reasons are never purely private. The alternative view is that while everyone has reasons of his own, they do not necessarily make claims on other people. An egoist thinks that other people have reason to benefit themselves but that these reasons do not necessarily provide him with reasons to do anything. It is a striking feature of normative reality that it is not this way. Other people and their reasons always *matter* for everyone, everywhere.

As noted, the Publicity Thesis has a certain initial plausibility. When I am in a position to interfere with you in the pursuit of your rational ends, I am always under some pressure not to do so. Of course, however, questions will arise. Suppose you have some reason to do something that it would be morally wrong for you to do — rob a bank, for instance. Does it follow that I have *some* reason not to interfere?

It is clear that I don't have a serious reason — the sort of reason it makes sense to dwell on. But as Schroeder (2007) has emphasized, it can be hard to tell the difference between a very weak reason, or an overridden reason, and no reason at all; so it can be hard to test claims like the Publicity Thesis by considering hypothetical cases, since apparent counterexamples can always be explained away.

Given this, it would be good to have an argument for the Publicity Thesis — a derivation of the thesis from deeper and independently plausible principles. And of course, even if we find the principle plausible and do not insist on an argument, we should still be in the market for an explanation for it. So we ask: What could possibly explain the truth of the Publicity Thesis?

Of course, it may just be a brute fact about normative reality that others' reasons always generate pressure not to interfere: an axiom in the theory of practical rationality. This would not entail that the Publicity Thesis could not be supported, for example, by appeal to coherence with our other views. But it would entail that there is no explanation for its truth.

Wallace is happy to endorse the Publicity Thesis simply on the basis of its reflective support and coherence with our other views. The Publicity Thesis, he contends, follows from "our best substantive understanding of what we (individually and severally) have reason to do" (2009: 471).[30]

While this reflective support is evidence for the truth of the Publicity Thesis, it would be disappointing if we could find no further reinforcement or explanation for it. The Publicity Thesis seems to be the sort of thesis that *cries out* for further explanation. We may all agree, as a substantive matter, that the reasons of other people bear on what we ought to do. But why should this be the case? Why *should* one take the reasons of others into account, rather than acting only for one's own sake?

Some philosophers argue that the Publicity Thesis does have a deep explanation (Korsgaard 1996, Bertea 2017). Unlike Wallace, they do not think that the Publicity Thesis is a substantive, "synthetic" truth, but rather that it follows from the very nature or concept of a reason — or an account what we are doing when we ascribe reasons to others. According to Korsgaard practical reasons are "public in their very essence" (Korsgaard 1996: 135). We

[30] Wallace argues that our normative judgments follow a pattern that supports the Publicity Thesis. Notably, "the judgment that S lacks good reason to do X seems necessary for our concluding that our interfering with S's X-ing is something we may be unproblematically be entitled to do" (485). By contrast, if we believe S to have reason to X, we take ourselves to be under at least some rational pressure to not interfere.

cannot coherently conceive of a reason that binds one person but not others. This sort of view is designed to answer the "why?" question in a decisive way, by providing a justification for the Publicity Thesis that should lead any reflective, truly rational skeptic to revise his view, irrespective of his substantive sensibility. Korsgaard offers a Wittgensteinian private-language argument for this conclusion which aims to show that we cannot coherently conceive of a reason that binds one person but not others. This argument, however, is widely rejected.[31]

Indeed, skepticism about the claim that reasons are "essentially" public is well-grounded. The skeptic who is happy to speak of reasons but thinks that your reasons have no claim whatsoever on me, does not seem confused about what reasons are or in violation of formal constraints on the ascription of reasons.

Nonetheless, I think there is reason to think this strong thesis is true. In this paper I offer a new explanation for the Publicity Thesis — one that, if true, entails that reasons are "essentially" public, though in a sense that differs somewhat from Korsgaard's (or Nagel's). The Publicity Thesis, I show, follows from an independently plausible form of non-cognitivism about normative judgment, and in particular about reasons judgments. If this form of non-cognitivism is true, then anyone who judges that R is a reason for X to Φ is rationally committed to judging that R is a reason to not interfere with X's Φing. Hence, everyone is committed to every instance of the Publicity Thesis by the nature of normative judgment.

The non-cognitivist view "entails" the Publicity Thesis indirectly — by entailing that everyone is committed to every instance of it. Reasons are "essentially" public in the sense that their publicity follows from an account of what we are doing when we ascribe reasons to others, rather than any substantive normative commitments.

Since the derivation is complicated, it's not surprising that the alternatives — e.g., the position of the egoist — do not *strike* us as confused. Nonetheless, if the non-cognitivist view is true, they are confused. One reason to accept the non-cognitivist view is that it explains the Publicity Thesis.

[31] Korsgaard's thesis has received extensive critical commentary. See Lebar (2001), Gert (2002), and Wallace (2009).

By way of preview, on the non-cognitivist view I defend the normative judgment that X should Φ consists in a kind of decision that X Φ. When the judgment is first-personal — e.g., my judgment that *I* should Φ — the view is similar to Gibbard's plan expressivism, though the state I call "decision" differs somewhat from a Gibbard-style plan. (I define a "decision" in functional terms, as a state whose essential role is to modify our intentions).

The deep difference between the views shows in the account of third-personal judgments. Gibbard endorses a view of third-personal judgments according to which they are "contingency plans" to do things in others' shoes; my judgment that (say) Mary should Φ is a plan on for me to Φ if I find myself in Mary's exact situation. On my view — the Simple View — the judgment as a decision for Mary *that Mary Φ*. The Publicity Thesis is a consequence of the Simple View and a corollary non-cognitivist view of judgments about reasons. In the first half of the paper, I will defend the Simple View over Gibbard's Subtle View. In the second half, I will show how the Publicity Thesis follows.

I will not defend non-cognitivism against alternatives, nor will I mount a complete defense of the version of non-cognitivism I endorse. I provide a more detailed justification elsewhere (Ayars 2021). This paper build on that one, focusing on the implications of the view of normative judgment for the Publicity Thesis. My aim in the first part is to show that my form of non-cognitivism is an improvement over the alternatives, and that it is plausible in its own right. The fact that Publicity Thesis follows from the account serves both as support for the account of normative judgment — since the Publicity Thesis is plausible — and as an explanation for the Publicity Thesis, a deep fact that we might otherwise be forced to regard as brute.

2. What is Normative Judgment?

The defense of the Publicity Thesis begins with an account of first-personal "should" judgments: the sort of all-things-considered judgments that close practical deliberation. These judgments purport to take all the reasons, moral and nonmoral, into account.

Suppose after my husband cheats on me, I conclude: "I should leave my marriage." What mental state does this judgment consist in? There are broadly two camps. One si*de se*es the judgment as an ordinary descriptive belief. The other sees the judgment as a non-cognitive attitude: a preference for leaving, an intention to leave, or some other conative state. Non-cognitivism has (over nonreductive cognitivist realism) the advantages of being consistent with a purely naturalistic metaphysics and delivering the tight internal connection between normative judgment and motivation that metaethicists have often emphasized.

The most developed and familiar form of non-cognitivism about "should" judgments is Gibbard's plan expressivism. According to Gibbard, to judge that I should Φ is to decide to Φ. The constitutive point of normative deliberation, says Gibbard, is not to represent an independent normative fact, but rather, to settle the practical question of what to do; hence normative judgment is a kind of decision. As Gibbard puts it: "Questions of what we ought to do are questions of what to do. Finish your deliberation, conclude what to do, and you've concluded what you ought to do" (Gibbard 2003: 19).

Let me make one important clarification about the mental state that we are calling "decision". This clarification will become important when I defend my distinctive view of third-personal judgments. Gibbard does not always distinguish the mental state of a "decision" or "plan" from an ordinary intention. But it is critical for the plan expressivist to insist on a distinction. (Since in ordinary language, to decide to Φ just is to form the intention to Φ, this involves a departure from ordinary language). If by "decision" we mean an ordinary intention, plan expressivism does not leave room for a paradigmatic form of irrationality: akrasia. An akratic agent judges that she should Φ without intending to Φ. If akrasia is possible — as it seems to be — the normative judgment expressed by "I should Φ" cannot be an intention to Φ.[32]

[32] The identification of normative judgments with ordinary intentions is also in tension with the fact that in those cases which clearly involve arbitrary picking, the resulting intention is not aptly expressed using normative language. Buridan's ass forms the intention to (say) go left, but this is not the judgment that he should go left rather than right. He may even have a general plan of always going left in such cases without thereby judging that he should go left. The only decision in our sense that Buridan's ass makes is the decision to go either right or

What is a "decision," then, if it is not an ordinary intention? Here is what, in my view, the plan expressivist ought to say: *A decision is the sort of state that concludes practical reasoning*. Decision is posterior to deliberation and just prior to intention. The akratic has concluded practical reasoning, and so made a "decision" in our sense — the same decision that a fully rational person would make. But decisions do not lead directly to action; they must be *executed* by forming an ordinary intention. The akratic's problem is at this stage, not at the stage of decision proper.

Crucially, practical reasoning concludes (and decision is thus complete) when the agent has taken his reasons into account in settling what to do, but before he has made the sort of arbitrary choice that Buridan's ass must make. The ass faces his practical problem and then reasons to the decision *to go either left or right* (but not to stand still), at which point there is no more practical *reasoning* to do. He then executes that decision by forming the intention to go (say) left; but that further step involves arbitrary picking, and so does not constitute a decision.

So far this is consistent with the cognitivist view that decisions are *beliefs about what should be done*, which lead to intention when all goes well. What distinguishes a decision from an ordinary descriptive belief is its functional role in our psychological economy. The well-functioning of a decision (and hence a normative judgment) does not consist in its representing an independent normative fact, as the cognitivist thinks. Rather its role is to settle or constrain our intentions "up to arbitrary picking" in a fashion responsive to our judgments about our reasons. Normative judgments are "supposed" to control our intentions in this way, and this is why they deserve to be called "decisions".

To take an analogy, suppose two people are on a road trip. The passenger, map in hand, is in charge of navigation. Her task is to instruct the driver where to go so as to promote her goal of arriving at her destination: "turn left here," "take the next exit," etc. The driver's role is

---

left. This disjunctive decision is properly executed by forming either an intention to go right or an intention to go left. But these intentions that execute the decision are not themselves decisions.

to execute the passenger's instructions, by setting the car in a certain direction. The passenger's decisions play the role of normative judgment in this example, and the driver plays the role of intention. There is clearly a reasonable sense in which the passenger is deciding where the car goes, even though she does not have control of the wheel — the role of her instructions is to lead to action on the part of the driver. Normative judgments are decisions in the sense in which the driver's instructions are: states whose role is to generate intentions.[33]

With this conception of decision in hand, we can take Gibbard's plan expressivism to be the view that normative judgments of the form *X should Φ* are constituted by decisions. In the case of first-personal judgments, this is plausible. When I deliberate and judge that I should leave my marriage, the judgment expresses the conclusion of my practical deliberation in a way that aims to take all of my reasons into account without making arbitrary choices. The action-guiding role of first-person normative judgment matches the action-guiding role of decision as we have defined it. Plan expressivism is designed to make sense of this close match.

3. Third-Personal Normative Judgments

However, we also make third-personal normative judgments: judgments about what other people should do. For instance, gossiping with you about Mary, I might say, "Mary should leave her marriage." What mental state does this third-personal judgments consist in for the plan expressivist?

I will distinguish two views: the Simple View and the Subtle View. The Simple View — the view I endorse — says that third-personal judgments are decisions *that* others do things. My judgment that Mary should leave her marriage is a decision for Mary that Mary leave, just as my judgement that I should leave is a decision for myself that I leave.

[33] Why does our mind have this dual-module structure, with the story of rational action proceeding in two steps: the decision and the intention? This is a hard question which I will not attempt to address here. But whatever the reason, this appears to be how we are wired. Our evidence for the claim that rational action proceeds in two stages is simply that we do make normative judgments prior to forming intentions.

Simple View: Just as my first-personal judgment that I should Φ is a decision that I Φ, my third-personal judgment that X should Φ is a decision that X Φ.

The Simple View is pleasingly straightforward, since it posits a simple correspondence between the content of the normative judgment and the content of the decisional state that allegedly underlies it. First-personal and third-personal judgments are represented as states of exactly the same kind, differing only in the content of the underlying attitude. But it is also *prima facie* implausible. First: what is a decision that *someone else* do something? On the functional definition of "decision" I just provided, a decision is a state whose essential role is to issue in an intention. So, a decision that Mary leave should be a state whose essential role is to issue in an intention on Mary's part. But Mary's actions are not under my control — my mental states cannot generate intentions in Mary. So how can my judgment that Mary leave be a state whose essential role is to regulate Mary's intentions?

Another worry is that the Simple View appears to misdescribe the motivational role of third-personal judgments. Decisions are desire-like attitudes in the following sense: a (first-personal) decision to Φ motivates the decider to Φ. So if third-personal judgments were decisions for others, it seems they would have to motivate the subject to promote their object. But you can judge that X should Φ without feeling the slightest motivation to promote X's Φing. (If Mary is my enemy, then my judgment that she should leave her marriage won't motivate me to help her leave — it may motivate me to sabotage her).

For these reasons, the Simple View is widely rejected. The main alternative —Gibbard's view — is that third-personal judgments are *conditional* first-personal attitudes (Gibbard 1983, 2003). They are contingency plans to do things in others' exact situations. My judgment that Mary should leave is a plan to leave my marriage if I find myself in Mary's "exact situation" — down to the last detail. Call this the

Subtle View: My first-personal judgment that I should Φ is a decision for myself that I Φ. My third-personal judgment that X should Φ is a decision for myself that *I Φ in a situation exactly like X's*.

Suppose Holmes is reasoning about what Mrs. Hudson should do. Gibbard says that Holmes cannot decide for Mrs. Hudson, but he can form a contingency plan for what to do in her situation:

> [I]f the two aren't in communication tonight [Holmes] knows that his own thinking can in no way settle the question of what to do in Mrs. Hudson's own mind. She can't possibly act on his thinking, the way Holmes later on can act on the planning he's now engaged in. Still, this is in a way a kind of hypothetical planning: he thinks what to do if in a hypothetical circumstance. He's thinking as if he could plan what to do if in Mrs. Hudson's plight … (Gibbard 2003: 50).

According to Gibbard, if Holmes judges that Mrs. Hudson should Φ, he is not only planning to Φ in Mrs. Hudson's exact qualitative situation, but planning to Φ if he is *Mrs. Hudson* in her qualitative situation. Third-personal judgments are plans for being other people. X's situation "down to the last detail" includes *being X*.

The idea that third-personal judgments are contingency plans for being other people in remote counterfactual situations may seem strange. We know we won't ever be another person — so these plans have no chance of being relevant. Yet, says Gibbard, despite the practical insignificance of this far-fetched contingency planning, we are nonetheless disposed to engage in it, and this is what we are doing when we make third-personal judgments about what others should do (Gibbard 2003: 53).

Despite this surface oddness, the view has a certain intuitive plausibility. After all, we do express our normative judgments by saying things like "I were you, that's what I would do." And it does avoid the problems for the Simple View. The Subtle View does not imply that third-personal judgments motivate the subject to promote their object; it only implies a conditional motivation, to Φ in X's situation. Moreover the essential role of my third-personal judgment that Mary should leave, according to the Subtle View, is to issue in an intention *on my part* to

leave in Mary's situation — not to issue in an intention on Mary's part to leave. And so it does not require the seemingly unfamiliar notion of "deciding for another person."

Yet, I think the Subtle View is unsatisfactory. In the next section I will present two objections to the Subtle View. I have raised these objections (plus an additional objection) already in (Ayars 2021), and I more detailed case against the Subtle View can be found there. The following is an abbreviated case against the Subtle View.

4. An (Abbreviated) Case Against the Subtle View

My first objection is that the Subtle View does not give a satisfactory account of interpersonal normative disagreement. Suppose you and I are gossiping about Mary. I judge that Mary should leave her marriage, and you judge that it is not the case that she should leave. Our normative judgments obviously disagree. They are inconsistent with one another. Thus, for the expressivist, the mental states that correspond to them must disagree.

On Gibbard's view, my judgment that Mary should leave is a plan that requires me to leave in Mary's situation. Your judgment that it's not the case that Mary should leave is a plan that permits you not to leave in that situation. If a single person made both these plans, his plans would be inconsistent, in the sense that one would forbid an action that the other permits. At least, this is what plan expressivists must say in order to account for why it would be inconsistent for a single agent to judge both that Mary should leave and that it's not the case that Mary should leave.

But a problem arises in the case of interpersonal disagreement. When I make the first plan and you make the second, my plan requires me to leave in Mary's shoes, whereas your plan says nothing about my conduct, and so neither forbids nor permits nor requires my leaving in Mary's shoes. My plan does not forbid any course of action that your plan permits; thus, the disagreement thus goes missing.

Of course there is a sense in which our plans "disagree" on the Subtle View. We make *de se* plans that would be inconsistent if a single person made them. But that's analogous to the sense in which we "disagree" if I judge that I'm in Princeton and you judge that you are

not. These judgments respectively self-ascribe properties that no single person could have, and therefore "disagree" in this sense. But of course they don't really disagree. If I say "I'm in Princeton" and you say "I'm not", though we are accepting and rejecting (respectively) the same centered proposition, we do not contradict one another in any sense. Gibbard's view entails that when I say "Mary should leave" and you say "No, she shouldn't", we disagree only in this sense — which is not any disagreement at all.

The second objection I will raise here is that the Subtle View conflates normative judgments that are distinct, in the sense that it is possible to hold one without holding the other (Gregory 2017[34]). The following judgments are given the same analysis on the Subtle View: My judgment that Mary should Φ and my judgment that I should Φ if I am Mary in Mary's circumstances. They are both construed as decisions to Φ in Mary's exact situation. But these judgments are distinct. Someone can accept the judgment she would express by saying "I should Φ if am X in X's exact situation" without accepting the judgment she would express by saying "X should Φ".

Consider an egomaniac with the following normative view: "In every possible situation, everyone should do what is best for me." The egomaniac accepts "if I'm Mary in Mary's exact situation, I (Mary) should do what's best for Mary." But he rejects "Mary should do what's best for Mary" since he does not believe that he is Mary.[35] (The egomaniac thinks that, since he's not Mary, Mary should do what's best for *him*).[36]

---

[34] Gregory presents a version of this objection (Gregory 2017: 43-44). Gregory's version uses the subjunctive conditional, rather than the indicative conditional. I am using the indicative conditional because it is clearer that indicative conditionals are analyzed as contingency plans in a Gibbardian framework. Indeed, Gibbard often states contingency plans in indicative form; for instance, he says that Brutus, in planning to stay home in Caesar's shoes, thinks "If I am Caesar in Caesar's shoes this morning, let me stay home" (Gibbard 2003: 6).

[35] The example may be more compelling if we imagine that the egomaniac does not know whether she is Mary. The indicative is infelicitous when the subject knows the antecedent is false. If the egomaniac does not know whether she is Mary, then she accepts the judgment expressed by "if I am Mary in Mary's circumstances, I should do what is best for me" but does not accept the judgment that Mary ought to do what is best for herself, since she does not know whether she is Mary. For more discussion of the conflation point and the egomaniac, see (Ayars & Rosen, forthcoming).

[36] Some philosophers think Qualitative Supervenience—the claim that there are no two possible situations alike in all non-normative qualitative respects but which differ in some normative respect—is a conceptual truth. If they're right, accepting one judgment but not the

The fact that the Subtle View conflates distinct judgments is a serious problem for it. It is a central desideratum for any non-cognitivist theory is that it provides distinct "acceptance states" for judgments that are clearly distinct, in the sense that it's metaphysically possible to hold one without the other.

5. The Simple View Solves the Problems for the Subtle View

The Simple View is not subject to these objections. On the Simple View, my third-personal judgment that X should Φ is a decision that X Φ, rather than a decision to Φ if I am X in X's situation. The problem of disagreement is avoided since, on the Simple View, the plans associated with "X should Φ" and "it's not the case that X should Φ " share an object: they both concern X and X's Φing. My judgment that Mary should leave is a plan that requires *Mary* to leave, and your judgment that it's not the case that Mary should leave is a plan that permits *Mary* not to leave. On the Simple View, disagreement between individuals is like disagreement within a single individual; since the contents in question are *de re/de dicto*, rather than *de se*, when individuals make plans with inconsistent contents they *ipso facto* disagree.

The Simple View also solves the conflation objection since it distinguishes the judgment expressed by "I should Φ if am X in X's exact situation" from the judgment expressed by "X should Φ". The first is a decision to Φ if am X in X's exact situation, the second is a decision for X that X Φ. The egomaniac makes a decision to do what is best for Mary if he is Mary in Mary's situation, without making a decision for Mary that Mary do what is best for herself.

The fact that the Simple View avoids these objections provides a strong *prima facie* case for the Simple View over the Subtle View. There are, of course, other reasons to favor the Simple View over the Subtle View. The Simple View gives a uniform account of first-personal and third-personal judgments, it makes third-personal judgments directly "about" other

other is a conceptual confusion. But what matters is whether accepting one judgment and not the other, is *possible*, not whether it is coherent; see Gregory (2017) for discussion. In addition, Gibbard denies Qualitative Supervenience, since he believes one can plan differently for being (say) Jane and for being Mary (Gibbard 2003: 50).

agents (rather than partly about oneself) and it resonates with a new and interesting moral psychology (which I will say more about in the next section). And, of course — as I will argue — it contributes to an explanation for the Publicity Thesis.

IV. The Simple View

Given the Simple View solves the problems for the Subtle View and has other advantages, it would be nice if the Simple View were viable. But to know whether it is viable, we need to say more about what "decisions for others" consist in and whether the objections to the Simple View stated earlier can be answered. In this section, I will further expound the Simple View show how the objections can be answered.

The Simple View says third-personal judgments are decisions for others. But what is a decision for another?

I will illustrate what I have in mind by means of a phenomenon I call *delegated decision-making*. Suppose that you and I are at a restaurant. You know more about what I would like from the items on the menu than I do. So I ask you to decide for me what I will order. If you agree to this request, you will deliberate on my behalf, taking account of all of the relevant factors (my preferences, my budget, and so on). And you will arrive at a conclusion which you can voice using normative language. Suppose you decide on my behalf that I am to get the burrito. Then your conclusion can be expressed either using an imperative — "Get the burrito" — or a normative assertion — "You should get the burrito."

When you say "you should get the burrito," you express a state which is aptly described as a decision for me. It is a decision for me in the sense that its role is to modify *my* intentions. Recall the functional definition of decision I offered earlier: a decision is a state whose role is to issue in an intention (in a way that responds only to the agent's reasons). So a third-personal decision — a decision for another — is a state whose role is to issue in an intention on the part of someone else. This is exactly the role of the state you express by saying "you should get the burrito." If all goes smoothly, it will lead me to form an intention to get the burrito, much as my own decisions/normative judgments issue in intentions on my part when all goes smoothy.

If I fail to form the intention, there is an interpersonal incoherence analogous to the intrapersonal incoherence involved in judging that one should Φ while failing to form the intention to Φ. Something has "gone wrong"; your state has not done what it is "supposed" to do.

Of course, the process by which your judgment leads to an intention on my part is indirect. You have to communicate your decision to me (unlike in the first-personal case, where no communication is necessary). The point is that the *role* your mental state plays is the same as the role my own decisions (normative judgments) play in my own psyche: to modify my intentions in light of my reasons. Simply because it is you who make the decision/judgment does not exclude it from playing this role. The claim is that the functional role of third-personal judgments extends "outside the head": they exert their characteristic effects, when they do, by affecting the thought and conduct of others.

In delegated decision-making, person A delegates a decision to person B who then *decides* (on behalf of A) *that* A is to do something. The fulfillment condition of B's decision is A's performing some action. There are two main sources of support for the claim that the judgments in delegated decision-making are decisions for others. The first is the role your state plays — as noted, its role is, intuitively, to influence my intentions. The second is that certain *rational norms* characteristic of decisions appear to govern these states.[37] Suppose the restaurant only has one burrito left. You know this, and you decide on my behalf that I am to get it. It seems that you cannot — once you've decided that I'm to get it — consistently decide to get it yourself. If after telling me I should get the last burrito you say to the waiter, "*I'll* get that last burrito," I would be puzzled, and would complain: "I thought you said *I* should get the burrito." (Similarly, if you say, "you should get the last burrito — and I should too," I would be puzzled).

[37] Roth argues on similar grounds that we can form intentions that others do things (Roth 2004, 2016). I am, however, making a claim about the state associated with normative judgment, which is not intention in the ordinary sense.

We can explain this by appeal to the fact that your third-personal judgment is a decision for me, together with the fact that the decisions of a given subject are subject to a co-realizability norm. If your judgment is a decision, it must harmonize with your other decisions, including your decisions for yourself, just as your decisions now about what to do at various times must be co-realizable. If you decide to Φ, and also to Ψ, it must be possible for you to both Φ and Ψ. And more generally, if you decide (in our technical sense) that X is to Φ and that Y is to Ψ, then it must be possible for X to Φ and Y to Ψ.

This *co-realizability constraint* derives from the action-guiding role of decisions. The role of a decision is to *settle* a practical question. Answers to practical questions must be compatible — otherwise, they cannot all give rise to action. In the first-personal case, this corresponds to the intuitively plausible *intrapersonal joint satisfiability* principle for "should": if X should Φ and X should Ψ, then it is possible that [X Φ and X Ψ]. In the third-personal case, it corresponds to the following *interpersonal joint satisfiability* principle: if X should Φ and Y should Ψ; then it is possible that [X Φ and Y Ψ]. If your decisions for others (and yourself) are not corealizable, then they cannot give rise to action (on the part of yourself and others) — and hence cannot fulfill their role. I will be saying more in defense of the interpersonal joint satisfiability principle later, when I address the Publicity Thesis.

The fact that the Subtle View *cannot* explain the felt incoherence of your judging that I should get the last burrito and then deciding to order it yourself counts in favor of the Simple View over the Subtle View. On the Subtle View of what your normative assertion expresses, there is no incoherence in saying, "You should get the last burrito", and then deciding to order the burrito yourself. Your judgment is simply a conditional plan to order the burrito if you are me in my exact situation, but — since you are not me — this conditional plan leaves you entirely free to decide get the last burrito in the situation you actually face. These decisions are co-realizable, since one is merely a conditional decision whose antecedent is not satisfied, so there should be no incoherence.

What I have said thus far answers the first objection: that there is no such thing as deciding for another. The example of delegated decision-making shows that we can and do

make bona fide decisions that refer to others' actions. It also shows that these decisions can be expressed using normative language, just as the Simple View predicts.[38]

Of course, my claim is that *all* third-personal "should" judgments (not just the states expressed in delegated decision-making) are decisions for others. One might be skeptical of this claim, even if one acknowledges that decisions for others occur in delegated decision-making. I will address this worry shortly.

But first, let me address the second objection to the Simple View: that it seems to misdescribe the motivational role of third-personal judgments. If third-personal judgments are decisions for others, and decisions are states which motivate the subject to promote their object, wouldn't my third-personal judgment that Mary should leave have to motivate me to help Mary leave?

The reply is that decisions are not in general states that motivate the subject to promote their object, and in particular that third-personal decisions do not imply such a motivation. On the present view, the function of a third-personal decision is not to motivate the subject of the decision (the decider) to act. Rather, it is to motivate the "executive subject" of the decision (Roth 2004) — the person whom the decision is a decision for — to act. So while your decision for me that I get the burrito is poised to generate an intention and a motivation in me (and likely will if it is expressed), you might not care at all whether I fulfill your decision, nor are you rationally committed to caring, and so need not be motivated to help me get the burrito. The decision is meant to settle a practical question *for me*, not for you. The Simple View is hence consistent with the phenomenological datum that third-personal judgments need not

[38] Some of the considerations I am invoking in support of this should be familiar. Philosophers have argued, on similar grounds, that we can form intentions which refer to other peoples' actions (Searle 1990; Gilbert 1990, 1996; Bratman 1992, 2013) and that the rational norms characteristic of intentions can extend across people (Roth 2004, 2014, 2016, 2018). I am, however, advancing a novel claim about the state associated with normative judgment. As noted earlier, for the plan expressivist normative judgment is not intention in the ordinary sense. My claim is that we can answer practical questions for others, and thus "decide for others" in the heavyweight sense of decision whose mark is aptness for expression in normative terms: e.g., "you should get the burrito".

be intrinsically motivating: their role is to influence others' intentions, not the intentions of the decider.

Isolating a case of a "decision for another" in delegated decision-making was helpful in seeing that decisions for others are possible, and why they do not imply motivations to promote the object of the judgment. But of course, my claim is that *all* third-personal "should" judgments — not just the states expressed in delegated decision-making — are decisions for others. Obviously, we make third-personal normative judgments outside the context of delegated decision-making, which is a "cooperative" social practice. We make third-personal judgments about strangers and enemies, and about people whom we know will not act on our judgments. Suppose that I privately judge that the president should end the war. This judgment has no chance of influencing the president's intentions — and I know this. The Simple View says that nevertheless, my judgment is a decision on my part that the president end the war. But is it plausible that normative judgments consist *across the board* in decisions for others?

First, let me be clear about the claim. In delegated decision-making the delegee enters into a normatively expressible state with a certain essential role: to influence the intentions of someone else. My claim is that whenever we make a third-personal judgment, we enter the same *underlying state*: a state whose role is to influence the intentions of someone else. It needn't be the case that the subject of the decision (the decider) care that the judgment fulfills the role — just as the subject of a belief need not care whether a belief is true. Still, what the judgment is "supposed" to do, in some sense, is to influence the intentions of someone else, just as a belief is "supposed" to represent the world.

The notion of the role or "point" of a mental state that I have been invoking is roughly the notion employed in the Gricean methodology of "creature construction" (Grice 1975; see also Bratman 2000). Creature construction involves imagining a series of increasingly sophisticated creatures which we can compare with actual creatures. For instance, the first creature in the sequence might have only desires and beliefs; the second might have intentions in addition to beliefs and desires; and so on. A mental state is characterized by what it enables

in a creature (and hence for us) — its role is its "point" in our psychological economy. Elucidating the role of a state then allows us to explain the rational norms on it. For instance, the norms of consistency and means-ends coherence on intentions are explained by the role of intention in planning agency (Bratman 1987, 2000).

The role of a state is therefore not just its causal functional role — it's typical inputs and outputs — but what it does for agents in the course of deliberate action. A characterization of the role of belief, for instance, would be something like this: the point of belief is to guide action by channeling intentions from ends to means in ways that benefit the agent. That is arguably the purpose for which beliefs as a class exist.

My claim is that the role of third-personal decisions/judgments — their raison d'être — is to modify the intentions of others, just as the raison d'être of first-personal decisions/judgments is to modify our own intentions. This introduces the idea that the relevant "creature construction" can invoke not only what a mental state does for a single agent, but also its "social" function — what it does for agents generally. The point of normative judgment is to influence intentions both intrasubjectively and intersubjectively, in light of reasons. We have the ability to stand back from our own impulses and ratify or not the course of action they seem to favor. But we also have the ability to guide the intentions of other agents by deliberating from their point of view and ratifying or not courses of action for them. We often do a better job of acting on our reasons if we step back and reason from them, bringing the full range of our background knowledge to bear. Similarly, it will often happen that others know more than we do and are less subject to some of the pathologies of decision than we are when we deliberate in our own case. And in these cases it can be valuable for others to deliberate on our behalf, to express those decisions, and for us at least sometimes to act on them. (Contrast this with Gibbard's view, according to which the usefulness of third-personal judgments is to "rehearse for life" by settling in advance what we will do in remote counterfactual scenarios (Gibbard 2003: 52)).

When the Simple View is understood as a claim about the role of third-personal judgments, the idea that normative judgments across the board consist in decisions for others

does not sound crazy. It is possible (and common) to enter a mental state even if it has no prospect — in a particular case — of fulfilling its normal output role. We said that the point of belief is to guide action by channeling intentions from ends to means in ways that benefit the agent. But of course, we form beliefs that have no chance of guiding action — e.g., the belief that the number of stars is odd or even, the belief that Julius Caesar spoke Latin. Similarly, though my judgment that the president should end the war has no chance of influencing the president's intentions, its official role could be to do just that. Thus, it is not implausible that my judgment that the president should end the war is a decision for the president that he end the war, where this is understood as a claim about the essential role of the mental state.

To sum up the argument so far: The Simple View avoids the objections the Subtle View (the problem of disagreement and the conflation objection). The two most pressing objections to the Simple View can be answered, as I have shown. Moreover, the claim that third-personal judgments are decisions for others is not as strange as it first appears. This is an excellent case for preferring the Simple View to Gibbard's Subtle View, and for taking it seriously as a better articulation of the plan expressivist's main idea.

V. Rational Constraints on Normative Judgment

I will now turn to main focus of this paper: the explanation of the Publicity Thesis.

An important part of the defense of the Publicity Thesis (as will become clear) invokes the rational constraints on normative judgments I mentioned earlier — in particular, the interpersonal joint satisfiability principle. This principle says that if X should Φ and Y should Ψ, then it is possible that [X Φ and Y Ψ]. One might be skeptical of this constraint. So I first need to show that this principle is independently plausible.

Recall the burrito example: If you decide on my behalf that I'm to get the last burrito, you cannot coherently decide to get it yourself. The Simple View explains this by identifying judgments with decisions and noting that decisions are subject to a co-realizability constraint: it is irrational to decide that A Φ and B Ψ if it is impossible that [A Φ and B Ψ]. The co-realizability constraint is what grounds the interpersonal joint satisfiability principle. It is in

virtue of the fact that third-personal "should" judgments are decisions for others, and that decisions must be co-realizable, that the interpersonal joint satisfiability principle is true of "should" judgments according to the present view.

The interpersonal joint satisfiability principle rules out certain combinations of normative judgments as incoherent, and has non-trivial consequences for the first-order views we can coherently accept. Of particular relevance to the defense of the Publicity Thesis is that the interpersonal joint satisfiability principle rules out judging both that X should Φ and that you should interfere with (in the sense of preventing) X's Φing.[39] The judgment that X should Φ is a decision that X Φ. The decision to interfere with X's Φing is a decision to do something incompatible with X's Φing. Since non-co-realizable decisions are incoherent, the corresponding normative judgments are likewise incoherent.

To take a more concrete example, suppose it would maximize your self-interest to acquire an object for yourself, but it would maximize my self-interest to interfere with your acquiring the object. On the Simple View, the following judgments are incoherent:

(J1) the judgment that you should take the object for yourself (because that's what would be best for you); and

(J2) the judgment that I should interfere with your taking the object for yourself (because that's what would be best for me).

On the Simple View, these judgments consist in the following decisions:

(D1) a decision for you that you take the object for yourself.

(D2) a decision for me that I interfere with your taking the object for yourself.

[39] I am using 'interfere' as a success term: if you interfere with X's Φing, x does not Φ.

But these decisions are incoherent in virtue of being non-co-realizable. The Simple View entails the incoherence of any normative view which prescribes incompatible courses of action for individuals.

But one might object to the Simple View on grounds that the rational constraints on normative judgment that it implies are implausible. The idea that normative judgments must be jointly satisfiable sounds right when the judgments are expressed and the context is cooperative, as in delegated decision-making and co-deliberation. But it seems to clash with commonsense in other contexts. Suppose I'm a burglar robbing your home and you catch me red-handed. I think to myself: "you should call the cops." But I know that cutting the phone lines would preclude you from doing this. Am I incoherent if I also judge that I should cut the phone lines? Intuitively not. So while the Simple View neatly explains the requirement of non-interference in delegated decision making, it appears to clash with intuition in other contexts.

In response, the proponent of the Simple View can say that the reason these judgments sound coherent is because we interpret them as judgments in the vicinity that are coherent. When I judge that you should call the cops and I should prevent you, perhaps what I mean is simply that it is in your self-interest to call the cops and it is in my self-interest to prevent you. Or — more likely —what I mean is that you should *try* to call the cops and I should try to cut the phone lines. These "trying" judgments are not incoherent.

What we often mean in contexts in which we say that people should do incompatible things is that each person should *try*.[40] Return to the case in which you can acquire the object for yourself, and I can interfere. We said that judgments that you should take it and I should prevent you are incoherent. But the following judgments are not incoherent on the Simple View:

(J1*) the judgment that you should try to take the object for yourself; and

(J2*) the judgment that I should try to interfere with your taking the object for yourself.

[40] See Dreier (1996: 415) for a statement of this idea.

since they correspond to the following co-realizable decisions:

(D1*) a decision for you that you try to take the object for yourself.

(D2*) a decision for me that I try to interfere with your taking the object for yourself.

D1 and D2 are co-realizable. If judgments like (J1) and (J2) strike us as coherent, this is plausibly because we are interpreting them as "trying" judgments.

Indeed, these observations may be turned into an argument for the Simple View. Note that on reflection, even in the absence of theory, there is something odd about (J1) and (J2) when they are not interpreted as "trying" claims. It would be fine to say that I should *try* to take the object for myself and you should try to thwart me. But it sounds odd to say outright that you *should* take the object for yourself and I *should* prevent you, since that's impossible.

Suppose Alice and Bob come to you for advice and ask what they should do if they find themselves in such a situation. It would be odd for you in your capacity as advisor to say: "Well, everyone should do what's best for himself, so you, Alice, should take the object for yourself, and you, Bob, should prevent Alice taking it for herself. That sort of advice sounds not just false but incoherent. This suggests that the interpersonal joint satisfiability may well be true. If it is, it is a point in favor of The Simple View that it entails it.

VI. Reasons Judgments and the Publicity Thesis

So far this is an account of all-things considered "should" judgments. But the Publicity Thesis is a thesis about reasons. To close the gap, we need an account of judgments about reasons. In this section, I will provide an account and show how the Publicity Thesis follows from the Simple View and this account of reasons judgments. Later, I will answer objections to the strong form of publicity implied by my non-cognitivist view.

The standard non-cognitivist view of judgments about reasons is that to judge that R is a reason to Φ is to weigh R as a consideration in favor of Φing. "Weighing a consideration in

favor" of an act is a psychological notion — it is something we do. When we are forming a judgment, are disposed to weigh certain considerations for and against various courses of action. For example, if you are trying to make up your mind about whether to stay home tonight or go out, you will weigh some considerations in favor of staying home and some considerations in favor of going out. Arriving at decisions by weighing considerations in this fashion is, as Gibbard says, something we could "program a robot to mimic" (2003: 190). Gibbard imagines a robot which first assigns weights to pairs of factors and actions, and then decides what to do based on this assignment of weights. Because judging that R is a reason to Φ involves weighing R in favor of Φing in deliberation about whether one *should* Φ (i.e., deciding to Φ), the weighing constitutes a kind of endorsement of R as counting in favor of Φing.

Weighing R in favor of Φing is something we do occurrently in deliberation about whether to Φ. But someone who is not currently deliberating may be in a state that disposes him to weigh R in favor of Φ when he deliberates. For this reason, judgments about reasons should be thought of — strictly speaking — as *dispositions* to weigh or *policies* of weighing considerations in favor of decisions (Bratman 2000). We can use the same phrase ("weighing in favor") for this underlying state; a person with a disposition to weigh R in favor of Φing "weighs R in favor" of Φing. To judge that R is a reason to Φ is to be in this underlying dispositional state.

So far this is an account of first-personal judgments about reasons: my own judgment about *my* reasons for Φing. But we also make judgments about the reasons of others, so we need an account of that. What is it to judge that R is a reason for *X* to Φ?

For Gibbard, to judge that R is a reason for Mary to Φ is to be disposed to weigh R in favor of Φing in Mary's exact situation (since judging that Mary should Φ is deciding to Φ in Mary's situation), but we've seen reason to reject Gibbard's Subtle View. Applying the "weighing considerations in favor" view of reasons judgments to the Simple View, we say instead that to judge that R is a reason for Mary to Φ is to be disposed to weigh R in favor of deciding that *Mary* Φ. Stated generally:

Simple View of Reasons Judgment: Just as my first-personal judgment that R is a reason to Φ consists in my weighing R in favor of deciding that I Φ, my third-personal judgment that R is a reason for X to Φ consists in weighing R in favor of deciding that X Φ.

But this natural adaption of the standard approach to judgments about reasons, when combined with the Simple View of should-judgments, puts us in a position to derive the Publicity Thesis; or better, to show that anyone who makes judgments about reasons is rationally committed to the Publicity Thesis. As we have seen, deciding that Mary Φ commits you to not deciding to interfere with Mary's Φing (given the interpersonal joint satisfiability principle). But this means that weighing R in favor of deciding that Mary Φ commits you to weighing R against interfering with Mary's Φing. Thus, a reason for Mary to Φ is also a reason for you: a reason against interfering with Mary's Φing.

The argument generalizes: for any R, any X and any Φ: anyone who judges that R is a reason for X to Φ is rationally committed to judging that R is a reason not to interfere with X's Φing. Hence, everyone is committed to every instance of the Publicity Thesis by the nature of normative judgment. Since no one can coherently reject any instance of the Publicity Thesis if this non-cognitivist view is true, the non-cognitivist view "entails" the Publicity Thesis by showing that no one can coherently reject any instance of it.[41]

[41] As Moore's paradox shows, we cannot always move from the observation that no one can coherently accept p to the conclusion that p is false, or not possibly true. There is a sense in which no one can coherently accept 'p but I don't believe that p', and yet propositions of this form are often true. But that is a very special phenomenon. In Moore cases, the incoherence only arises if the claim is asserted or consciously and reflectively tokened. If belief can be utterly unconscious, then there is nothing incoherent in believing p (unconsciously), and also <I don't believe that p> consciously or otherwise. The incoherence at issue in the present case does not depend on overt assertion or conscious endorsement. It is incoherent to have non-corealizable plans whether those plans are conscious or not; attitudes incompatible with the Publicity Thesis are similarly incoherent in every case. That gives the sense in which the present view "entails" Publicity.

To get the intuitive idea, recall the example in which I ask you to decide for me what to get at the restaurant. In this context, if you judge that R is a reason for me to get the last burrito, you will also weigh R against deciding to get it yourself, since you are deciding for both of us based on my reasons as well as yours. Reasons take on an intersubjective dimension: a reason for you (to do one thing) is a reason for me (against preventing you from doing it). If the non-cognitivist view I am defending is true, this holds in full generality. Even if there has been no explicit delegation of authority, I am rationally required to take your reasons into account in making decisions for myself, since in making judgments about your reasons I am weighing considerations in favor of decisions that I will make for you, which must harmonize in the end with my decisions for myself.

Strictly speaking, if this defense goes through, what it shows is that no one can consider and reject an instance of the Publicity Thesis. *If I judge that R is a reason for you to do something,* I must also judge that R is a reason for me not to interfere. But what if I'm not focused on what you should do, and so make no judgments about it? What if I am only focused on myself, and my own reasons, and never consider what your reasons are?

The blinkered agent who never thinks about other people is not yet under any pressure to think about them, and to constrain his judgment about his own reasons by his judgments about theirs — at least, so far as the claims I have made so far are concerned. The argument I have made therefore cannot reach a person to whom it never occurs that others have reasons of their own. So I have not shown that we are all under rational pressure not to interfere with others. The rational pressure comes from wide-scope constraints on combinations of normative judgments. Someone who abstains from normative judgments about others will not feel this pressure, just as someone who does not accept the premises of a valid argument will feel no wide-scope pressure from norms of consistency to accept its conclusion.

Yet it is still highly significant to show that anyone *who makes judgments about the reasons of others* is committed to constraining his judgments about his own reasons by his judgments about theirs. This principle rules out the egoistic view discussed at the start. The egoist acknowledges that other people ought to do things and have reasons of their own for

doing them but thinks these reasons do not necessarily make a claim on him. The present view shows this outlook to be incoherent. Again, what the view entails that no one can coherently reject *any instance* of the Publicity Thesis: for any R and any Φ, it is incoherent to judge that R is a reason for X to Φ while rejecting that R is a reason against interfering with X's Φing.

Of course, the Publicity Thesis is a merely formal thesis. It says that *if* R is a reason for X to Φ, then R is a reason for anyone to not interfere with X's Φing. It places no constraints on what R may be. It does not tell us, as substantive matter, what our reasons (or the reasons of others) are. An egomaniac, who thinks that everyone has reason only to benefit him, may respect the Publicity Thesis (insofar as he also judges that these reasons give him reasons not to prevent others' from benefiting him). Of course, the egomaniac's view is not altruistic in any sense. So the non-cognitivist view should not be thought of as entailing altruism. It entails only a formal thesis about reasons. But the formal thesis is still highly significant. It means that the only possible reasons are reasons that we share.

In fact the conclusion is stronger. It is not just that your reasons to Φ have *some* weight for me as reasons not to interfere with your Φing. On the present view, a reason for you is a reason *of equal force* for me not to interfere. The stronger your reason to Φ, the stronger my reason not to do anything incompatible with your Φing. The significance of others' reasons, on the present view, is analogous to the significance of your own reasons in generating rational pressure not to interfere with your own rational actions. If you weigh R (as a reason of strength s) in favor of deciding to Φ on your own behalf, you are committed to weighing R (as a reason of strength s) against doing anything incompatible with Φing, since deciding to Φ commits you to not deciding to do anything incompatible with Φing. This corresponds to the intuitive principle that if R is a reason to Φ, R is a reason of equal force against doing things incompatible with Φing. The analogous principle, on the present view, is that if R is a reason for X to Φ, R is a reason of equal force against doing things incompatible with X's Φing.

This form of publicity is relatively strong — so strong, in fact, that it raises worries about the intuitive plausibility of the view. In the next section, I will answer objections to this strong form of publicity.

VII. Publicity and Partiality

When I am in a position to interfere with you in the pursuit of (what I judge to be) your rational ends, I am always under some pressure not to do so. If you have reason to feed your dog or build a spaceship and I can stop you, the mere fact that you have your reasons means that I need *some* fairly strong reason to interfere; to interfere on a whim would be unreasonable; that much is intuitive. But if the non-cognitivist view I am defending is true, I am under *considerable* pressure. The strength of my reason not to interfere is not just proportionate to (what I take to be) the strength of your reason to perform the act; the strengths are equal. And it may seem implausible that others' reasons have this degree of significance. This seems to imply that our own interests (or the interests of those near and dear) give us no greater reason than the interests of strangers in certain cases. And that can seem hard to believe.

Recall the shipwreck case: you and I are fighting for the last plank, but it is built into the setup that if I save myself, I will thereby prevent you from saving yourself. Call cases of this sort — cases in which it is not possible for everyone do what we pre-theoretically think he has most reason to do — *conflict cases*. The non-cognitivist view entails that in conflict cases, it's not the case that each of us *ought* to get the plank for himself. But we can still ask whether each of us has *reason* to get the plank for himself, and the answer to that question might be "yes." But on the present view, a reason for you to Φ is a reason of equal force for me not to prevent you from Φing. So, your weighty reason to get the plank for yourself translates to an equally weighty reasons for me not to prevent your getting the plank for yourself. That seems to imply that in these conflict cases, my reasons to grab the plank for myself are no stronger than my reasons to cede the plank to you. But this looks wrong. It seems you have more reason to take the plank for yourself than to cede it.

It looks especially wrong in cases like the following. Suppose you and I are fighting for a bottle of medicine needed to save the life of our respective children. It is highly intuitive to

think that in this case, we each have more reason to take the medicine for our own child than to cede it to the other. If you can overpower me and grab the medicine, you should do so — or so it seems. This is the case even if my child needs the medicine slightly more than your child does.[42] We are not rationally required to be totally impartial.

How can the present view accommodate this? Again, the problem is that, if the non-cognitivist view is correct, any reason that you have to take the medicine generates a reason of equal force for me not to prevent your taking the medicine. Since my taking the medicine entails that I prevent you from taking it, these reasons not to prevent your taking it counterbalance my reasons to take it. So, in conflict cases, each agent has no greater reason to take the medicine for herself than to cede it.

I take this to be hardest challenge to my view, and I see no way to avoid this implication. I accept that my account implies that in the medicine case each agent has no greater reason to take the medicine for herself than to cede it. But, I will argue, this result is not as problematic as it may first appear. The apparently counterintuitive implications can be softened by observing that the non-cognitivist view is consistent with the intuitively correct verdict about what each agent *should do* in cases like the medicine case: namely, that each agent should try to take the medicine for herself. I will explain why.

The first thing to note is that even in conflict cases, each agent's reasons do generate strong reasons of non-interference for the other. The fact that the medicine is needed to save your child's life provides an extremely weighty reason against my interference. Consider that if I were not also in the desperate situation of needing the medicine for my own child (or something analogous) I clearly should not interfere with your taking it. In the plank case, the reasons to cede the plank are so strong that, arguably, it would not be unreasonable to cede the plank (though it would also be reasonable to try to take it). This is something that needs to be explained. The non-cognitivist view I am defending explains it, and this is a serious advantage of the view.

[42] Thanks to Johann Frick for this objection.

The problem is that the non-cognitivist view is inconsistent with the intuitive thought that, nonetheless, these reasons are outweighed in at least some cases of conflict. The commonsense view appears to be that though I have strong reason to not interfere with your getting the medicine, I have stronger reason to take the medicine myself.

The non-cognitivist view is inconsistent with this verdict, but this result is not as problematic as it first appears. The proponent of the non-cognitivist theory can respond by noting that the theory can accommodate the view that what each agent *should do* in cases like the medicine case is try to take the medicine for herself. This is because the non-cognitivist view is consistent with the (plausible) claim that each agent has reasons to try that are not counterbalanced by reasons against trying. Reasons to try to Φ are special because, while they translate to reasons for everyone not to interfere with the agent's trying to Φ, these reasons do not weigh against others' trying to Φ (or Φing) themselves.

The critical feature is that trying to take the medicine for myself does not interfere with your *trying* to take it for yourself. Though I may be able to prevent you from taking the medicine, thus interfering with your *getting* it, I cannot prevent you from trying to take it; even if I manage to grab it for myself, you will still have tried, so I will not have prevented you from trying. So, though your reasons to try, per the present theory, generate reasons for me not to interfere with your trying, these reasons have no practical force, since I can't possibly interfere with your trying (in the sense of preventing it). In particular, they do not generate reasons against *my* trying to Φ (or my Φing). Simply put: my trying does not interfere with your trying; indeed, my succeeding does not interfere with your trying.

So, here is how we explain how each agent can be rationally required to try to benefit himself in conflict cases. In the medicine case, we say our respective reasons to try to get the medicine for our respective children tip the balance in favor of each of us attempting to snatch the medicine before the other does. (I will say more about what these reasons might be shortly). The verdict that each of us should try is first of all coherent, since it is possible for

both of us to try, and second correct, since — I will argue — both agents have non-counterbalanced reasons to try to Φ.[43]

Of course these reasons to try to get the medicine must be distinct from the reasons each of us has *to get the* medicine, since according to the theory, reasons of this latter sort are always exactly counterbalanced by reasons not to interfere with the other's getting the medicine. So the proposal requires that there be reasons to try to Φ that do not derive their force from prior reasons to Φ. But are there really reasons to try to Φ that are distinct from reasons to Φ? I will suggest that there are, and that it is plausible that there are reasons of this sort for each agent to try in the medicine case.

The most obvious examples of reasons to try are in cases of ordinary, friendly competition, as when you and I are playing a game of Monopoly or table tennis. To count as playing the game at all, each of us must try to win. If we have reason to play, we have reason to try. The whole fun of it lies in the trying. Insofar as we praise people for traits like grit and perseverance, we recognize a value to trying.

In these cases, the non-cognitivist view predicts the pattern of judgments that seem correct. It is not the case that each should win (since that is incoherent). But it is the case that each should try to win, since each has non-counterbalanced reasons to try. So a neutral third party deciding for everyone can and should decide that each should try. (This is what we want to say in the medicine case, though in that case the reasons to try must have a different source).

It may be that reasons to try to Φ are conditional on Φing being an option for the agent; trying to Φ has little or no value if Φing is not a possibility (and may be impossible if the agent *knows* that Φing is not a possibility). But this does not mean that all reasons to try are derived from reasons to Φ. When an agent exhibits grit or perseverance, we praise her because the trying itself exhibited a kind of virtue, which is not reducible to the value of trying in bringing about her Φing — even if Φing must remain possible for the agent for trying to Φ to be virtuous

[43] When one agent can overpower the other, their trying to Φ entails their Φing — acting on their reasons to try entails Φing. The claim is that it is rational for the agent to Φ not because she has greater reason to Φ than to not Φ, but because acting on her reasons to try to Φ entails her Φing.

in this way. And similarly in cases of friendly competition: even if it's true that you have reason to try to win only when it's in some sense possible for you to win, still your reasons to try do not depend on your reasons to win. Trying in this context can be fun even if winning would not be especially fun, and it's the prospect of fun that generates the reasons to try.

It is plausible that we also have reason to try in the medicine case. The special relationship we have to our respective children entails an obligation to *try* to do what is best for them. Why might we have such reasons to try? The action of trying is expressive — of love, loyalty, and so on. The fact that your child is *yours* gives you reason to do things that express your love and loyalty to your child. This includes trying to save your child's life, even when it precludes me from saving the life of my child.

The critical feature of trying, as noted, is that your trying does not interfere with my trying in a way that expresses my love for and loyalty to my child. This means that your relationship-based reasons to try are not counterbalanced by reasons not to interfere, and are therefore fit to tip the balance decisively in favor of your trying to benefit your child. And the same goes for me. So in this sort of case the theory predicts exactly what common sense maintains: that in desperate conflict of this sort, each of us has (decisive) reason to struggle for the medicine. If I were to note the symmetry and decide to (say) flip a coin — heads I struggle, tails I back off and cede the medicine to you — I would not be doing what I have most reason to do. To do that would be to act against my weighty and non-counterbalanced relationship-based reasons to save my child.

What about conflict cases involving competition in narrow self-interest, as when you and I are fighting for the last plank? It is not plausible that we have reasons to try in this case deriving from an expressive function of trying. Yet in this case, the theory arguably predicts the intuitively correct result about what each of us should do. The theory predicts (assuming there are no independent reasons to try) that it would be rationally permissible to try, and also rationally permissible to cede, since the reasons to take the plank are perfectly counterbalanced by the reasons to cede it. And this seems to be the commonsense view. Commonsense says that while it would not be unreasonable for me to try to take the plank for

myself, it would also not be unreasonable for me to let you have the plank. (We might praise a person who did this as selfless).

So — if the commonsense view is that either option would be reasonable — the non-cognitivist theory is not counterintuitive. It's only if we think that it would be in some way unreasonable simply to cede the plank that the theory pinches.

Perhaps we do think it would be unreasonable: each ought not cede the plank to the other. But if we take that view, we can say that this is because there are plausibly other factors that ground reasons to try in the shipwreck case. Here is one hypothesis: we each have reasons to try deriving from the value of self-respect. It is intuitively plausible that self-respect requires each of us not to be self-abnegating. Someone who always cedes to others is self-abnegating in a way that is incompatible with a strong form of self-respect. Thus, if I were to (say) flip a coin, rather than trying to acquire the plank for myself, I would not be doing what I have most reason to do, because I would be failing to respond to my reasons of self-respect to try to take the plank for myself.

To sum: We started with observation that in conflict cases, it seems that each agent has greater reason to (say) take the medicine than to cede it. The proponent of my non-cognitivist view must deny this, since in conflict cases, each agent's reasons to take the medicine are counterbalanced by reasons to not prevent the other from taking the medicine. However, this result is not as problematic as it may first appear. The apparently counterintuitive implication can be softened by observing that the non-cognitivist view is consistent with the intuitively correct verdict about what each agent *should do* in cases like the medicine case: namely, that each agent should try to take the medicine for herself. This verdict is possible because each agent has reasons to try that are not counterbalanced by reasons not to try. The reason these reasons are not counterbalanced is that though each agent's reasons to try, per the present theory, generate reasons for the other not to interfere with the other's trying, these reasons have no practical force, since each agent cannot possibly interfere with the other's trying (in the sense of preventing it).

In fact, it is not obvious that the commonsense view is that each agent has greater reason to *take* the medicine than to cede it. This alleged intuition may really be the intuition that each agent has reasons to *try* to take the medicine that are not outweighed. Consider that in cases in which there are plausibly no reasons to try — as in the shipwreck case — our intuition is (arguably) that it would not be unreasonable to cede the plank. This suggests that our judgment that it would be unreasonable to cede in the medicine case is produced by the thought that there are strong reasons to try in the medicine case deriving from the expressive value of trying — reasons to try which are not present in the shipwreck case. If so, the non-cognitivist view predicts the perfectly predicts the pattern of judgments that seem correct, including the pattern of judgments about the *reasons*. This would be a strong point in favor of the non-cognitivist view.

VIII. Conclusion

In this paper I explained a significant thesis about reasons from an independently motivated non-cognitivist account of normative judgment, in particular, "ought" judgments and reasons judgments. If this form of non-cognitivism is true, then anyone who judges that R is a reason for X to Φ is rationally committed to judging that R is a reason to not interfere with X's Φing. Hence, everyone is committed to every instance of the Publicity Thesis by the nature of normative judgment.

BLAMING FOR UNREASONABLENESS:
Accountability without Ill Will

(Published in *Journal of Ethics and Social Philosophy*)

https://heinonline.org/HOL/Page?handle=hein.journals/jetshy19&div=6&g_sent=1&casa_token=

ABSTRACT

Many theorists of moral responsibility endorse the following quality of will condition: *X is blameworthy for a wrong act A only if in doing A, X expressed ill will.* I argue that the quality of will condition is in fact false to ordinary moral practice, because ordinary practice licenses blame for agents who act wrongly from epistemically unreasonable ignorance of the wrong-making features of their act, even when the act does not express ill will. In response, I provide a theory of culpable ignorance, and blameworthiness more generally, on which ill will is not necessary for blameworthiness. My theory, which I will call *Rational Capacitarianism*, entails that when an agent acts wrongly from factual ignorance, she is blameworthy if and only if her ignorance is caused by a failure to exercise a *rational capacity*—a capacity to recognize reasons for believing and desiring and to be generally responsive in one's beliefs and desires to one's assessment of reasons. Failures of rational capacities are relevant because only such failures can be attributed to the agent in a distinctive way. My theory shows that we do not need to choose between a quality of will view (which seems too lenient) or a standard Capacitarian view (which seems too harsh).

1. Introduction

Many theorists of moral responsibility endorse the following:

> **Quality of Will Condition**: *X is blameworthy for a wrong act A only if in doing A, X expressed ill will*

Ill will is, roughly, an objectionable lack of concern for morality or the morally significant interests of others. Exactly what ill will consists in is controversial. All hands agree that ill will need not involve malice, so the term is something of a misnomer: it would be better to speak of an "insufficiently good will" or perhaps an "objectionable pattern of concern." However, nothing I say in this paper turns on any particular view of ill will, other than that ill will is a matter of what the agent wants or cares about.

On one prominent conception, an agent's quality of will is a matter of her responsiveness to moral reasons (Arpaly 2002). A good willed person has a final desire to take courses of action that have right-making features (like kindness or fairness), and a final desire to avoid actions with wrong-making features (like cruelty). In Arpaly's words: "To say that a person acts out of moral concern is to say that a person acts out of an intrinsic (noninstrumental) desire to follow (that which in fact is) morality" (Arpaly 2003: 84). And a person acts from ill will if she is either insufficiently responsive to moral reasons, or responsive to sinister reasons—"reasons which conflict with morality" (Arpaly 2003: 79). A person who tortures a puppy for fun expresses ill will because she desires the puppy's pain, and is thus responsive to a consideration that makes the act wrong. A person who fails to aid someone who is suffering might express ill will, not because she is responsive to a sinister reason, but because she is insufficiently responsive to the moral consideration at hand—the person's suffering.

Proponents of the quality of will condition sometimes argue that it is implicit in our ordinary practice of blaming. Strawson (1962) famously claimed that the *reactive attitudes* —

resentment, indignation, gratitude, and the like — are always responses to the good or ill will displayed by others:

> The personal reactive attitudes rest on, and reflect, an expectation of, and demand for, the manifestation of a certain degree of goodwill or regard on the part of other human beings towards ourselves; or at least on the expectation of, and demand for, an absence of the manifestation of active ill will or indifferent disregard (1962: 347).

Indeed, the quality of will condition appears to explain central features of our practice, most notably the range of excuses we accept as grounds for withholding blame for wrongful action. It is a central feature of our blaming practice that we withhold blame when we take the agent to have acted under duress or from certain forms of blameless ignorance. Any theory of blameworthiness must explain why these excuses work when they do, and for the Strawsonian the explanation is straightforward: these excuses negate blameworthiness (when they do) by indicating that the violation was not sourced in any improper concern for the rights or interests of others. It is important to note that the quality of will condition only specifies that ill will is a *necessary* condition of blameworthiness. It is consistent with thinking there are other requirements, e.g., that the agent acted wrongly, or that she possessed the capacity to recognize and respond to moral reasons. Proponents of the quality of will condition include Strawson, Arpaly (2003), Rosen (2015), Wallace (1994), McKenna (2012), and Graham (2014).

I will, first, argue, against this near consensus, that the quality of will condition is in fact false to ordinary moral practice.[44] This is because ordinary practice licenses blame for agents who act wrongly from epistemically unreasonable ignorance of the wrong-making features of their act, even when the act does not express ill will. This should be especially concerning to Strawsonians, who think the norms implicit in our practice are the inevitable starting point for discerning the real norms of blameworthiness.

[44] Baron (2014) is a fellow traveler in this position.

Second, I provide a theory of culpable ignorance (and blameworthiness more generally) on which ill will is not necessary for blameworthiness. The theory is an answer to the question: in what might the culpability of unreasonable agents be grounded, if it is not ill will? There is already an answer to this question in the literature: Capacitarianism, which says that an agent is culpable for fully unwitting wrongdoing insofar as she had, but failed to exercise, the *capacity* to be aware of the wrongness or wrong-making features of her act. But all Capacitarians need to distinguish between capacities that are relevant to responsibility and capacities that are not, as will become clear. The present approach draws this line in a principled way.

My theory, which I will call *Rational Capacitarianism*, entails that when an agent acts wrongly from factual ignorance, she is blameworthy only if her ignorance is caused by a failure to exercise a *rational capacity*—a capacity to recognize reasons for believing and desiring and to be generally responsive in one's beliefs and desires to one's assessment of reasons. I argue that failures of rational capacities are relevant because only such failures can be attributed to the agent in a distinctive way. My theory shows that we do not need to choose between a quality of will view (which seems too lenient) or a standard Capacitarian view (which seems too harsh).

A few preliminaries: I will consider an agent to be blameworthy for an act if she is liable to certain "hot" emotional responses for it, like resentment and indignation. This follows Strawson in reducing questions of blameworthiness to questions about the appropriateness of the negative reactive attitudes.

I will deviate from the literature in focusing on ignorance that results from epistemic unreasonableness, rather than lapses of memory or inattention. The literature on negligence tends to spotlight forgetful or inattentive agents, e.g., a parent who forgets about a child in a hot car, or a driver who does not check his rear-view mirror and backs into a pedestrian. But I think it is a mistake for ill will skeptics to emphasize these cases, because it is actually *less* apparent that forgetful or inattentive agents are culpable than it is that *unreasonable* agents are culpable. This will become clearer when I explicate the notion of a rational capacity.

2. Epistemically Unreasonable Ignorance

There are two relevant kinds of ignorance: factual and moral. A factually ignorant agent is unaware that her act has certain empirical features that make it wrong (like being risky or harmful); a morally ignorant agent may be aware of these features, but ignorant that her act is wrong (or cruel or unfair). I will focus on factual ignorance rather than moral ignorance, because it is more difficult to see how an agent whose pertinent ignorance derives entirely from factual ignorance might be culpable.[45] Factual ignorance is often a complete excuse. Suppose X poisons Y by putting a substance in her coffee. If X innocently believed that the substance was sugar (perhaps because it was mislabeled, and she had no reason to doubt the labeling), then X is excused—she is not blameworthy for the wrong she committed.[46] Strawsonians explain this in a straightforward way: if X was ignorant of the harm she would cause and acted only because she was ignorant in this way, the wrong was not sourced in any improper concern.

Factual ignorance is only an excuse, of course, when it is not the foreseeable upshot of some prior blameworthy act which expressed improper concern (Rosen 2003). If to relieve the boredom of a long drive I ingest a mind-altering drug and hit a pedestrian whom I don't notice crossing the street, it is no excuse that I didn't know he was there. Since I am blameworthy for recklessly taking the drug, I am blameworthy for my ignorance.

---

[45] There is an obvious way in which morally ignorant agents can still express ill will, by being insufficiently responsive to the reasons that *make* their act wrong. A Nazi who tortures innocent prisoners fully convinced he is doing the right thing is still insensitive to the suffering of the prisoners. This is why many quality of will theorists do not think *moral* ignorance is exculpating.

[46] Harman (2011) points out that it is not ignorance that is potentially exculpating, but false belief. If you have a .5 credence that a substance is poison and a .5 credence it is sugar, then you're ignorant that it is poison, but this doesn't get you off the hook for poising someone. You're only off the hook if you have the false belief that the substance is *not* poison. This point is well-taken. "Ignorance" of p in this paper should be construed as shorthand for false belief that not-p; further qualifications may also be needed (see Harman 2011 for discussion).

Ignorance may also not excuse if it is *motivated* by ill will, as when a racist's racial animus leads him to falsely believe a black candidate is unqualified for a job. If an agent acts from motivated ignorance, she acts from improper concern, since her ignorance expresses improper concern. Call ignorance that satisfies neither of these conditions — that is neither motivated by ill will nor the upshot of prior blameworthy conduct —*morally blameless* ignorance (leaving open that it may be criticizable in some other way, e.g., epistemically criticizable). In what follows I set aside morally blameworthy ignorance. The plan is to focus on *blameless ignorance* and to ask when and why it amounts to an excuse.

Most theorists hold that acts performed from blameless factual ignorance are always blameless. For instance, Rosen says:

> If [the agent] has been neither negligent nor reckless in the management of his opinion, then his ignorance is blameless and so is the act done from ignorance (Rosen 2003: 63)[47,48]

And Zimmerman says, "culpability for ignorant behavior must be rooted in culpability that involves no ignorance" (Zimmerman 1997: 417). Indeed, it may seem almost a truism that an agent is not blameworthy for a wrong act if she was blamelessly unaware of its wrong-making features. Consider this case of Arpaly's (2002): Boko Fittleworth (a character in a P.G. Wodehouse novel) captures a man he finds hiding in his future father-in-law's garden shed, thinking he is a burglar. In fact, he is not a burglar; his presence is part of a secret plot in which the future in-law is willingly participating. Boko is, obviously, completely blameless for this honest mistake.

But Rosen's view is not as obvious as it may first appear. We must distinguish *epistemically reasonable* from *epistemically unreasonable* ignorance. The cases that support the

[47] By "neither negligent nor reckless," Rosen means that the person's ignorance is not the upshot of a previous negligent or reckless *act*. It does not cover pure epistemic irrationality, which will be my focus.

[48] Rosen's view is actually more extreme, as he thinks that blameless moral ignorance is also always excusing. But we can set this aside.

view that blameless ignorance is always exculpating tend be cases of epistemically justified ignorance: cases in which the agent drew sound conclusions based on the evidence available to her. I am clearly off the hook for poisoning you if my belief that the substance was sugar was not only unsourced in ill will, but completely reasonable (for instance, because the canister was labeled "sugar" and I had no reason to think this was misleading). But what if the false belief from which an agent acts is, though blameless, highly irrational? Suppose the canister was labeled "cyanide—not for consumption!" but I unreasonably, though innocently, assumed it was a joke. Am I still off the hook?

Now, irrationality is sometimes the result of underlying bad will. I wish to set these cases aside (since they are instances of *motivated* ignorance), but since it may be hard to distinguish them from cases of innocent irrationality, it is worth saying a bit more about them. In *Unprincipled Virtue*, Arpaly asks us to consider someone who claims "I hate him because he is disgusting." We are likely to infer that the speaker "puts the cart before the horse" (Arpaly 2003: 105): it is the hate that motivates the belief that the person is disgusting, not vice versa. Consider a young academic who irrationally believes that women are awful at abstract thought, despite being surrounded by brilliant female colleagues. Again, we would reasonably surmise that the man's irrational belief is rooted in some unsavory motive, perhaps a discomfort with women's presence in academia. These cases of *motivated* irrationality or *motivated* ignorance are clearly possible. And when an agent acts from motivated ignorance, she acts from ill will in a straightforward sense, because her ignorance itself is the result of a bad desire.

The cases that interest me are not like this. They are ones in which the agent's irrationality is unmotivated. It is clear that human beings are sometimes just poor thinkers. As Kahneman and Tversky (1986) have shown, we are subject to a range of cognitive biases. For instance, we tend to be bad at reasoning about probabilities because we neglect base rates and rely too heavily on examples that immediately come to mind (the "availability heuristic"). And sometimes, of course,

people simply form beliefs that are incongruent with the overall balance of their evidence, even when they are not affected by a particular bias.

Given the prevalence of garden-variety irrationality, it is clearly *possible* that someone might act badly from ignorance that reflects unmotivated irrationality but which is not further sourced in some improper pattern of concern. Whether or not such cases are common is, of course, another matter. What I hope the reader will appreciate at this point is that such cases are *possible*, and that the right theory of blameworthiness should properly classify them. Are agents who act wrongly from *unmotivated* irrational ignorance culpable?

Of course, it can be very hard to tell in any given case whether the agent's irrational ignorance is badly motivated or not. That's why, in constructing cases for the purposes of inspecting our sense of culpability, we should make it as clear and vivid as possible that the agent's pattern of concern is entirely appropriate. But some help from the reader can't hurt: I ask the reader to imagine the characters as being as good-willed as they could possibly be—as the sort of people who really are concerned with morality and others' interests as much as they should be, but who suffer from bouts of ordinary unmotivated irrationality.

3. Some Cases

In *Who Knew? Responsibility Without Awareness*, Sher discusses cases of "poor judgment" for which we are likely to blame the agent (Sher 2009: 26). In *Home for the Holidays*, Joilet, a homeowner afraid of burglars, hears movement in the kitchen. She grabs her gun, runs downstairs, and shoots the intruder, only to discover that it was her son who came home early for the holidays. In *Colicky Baby*, Scout gives a crying baby a mixture of vodka and juice to ease its digestive pains.

In each of these cases, it is possible — in fact, easy — to interpret the agent as acting from improper concern. We suspect that Scout knows, at least at some level, that giving a baby vodka is dangerous and just doesn't care enough. But it is also *possible* that the agent is properly

concerned but merely foolish, which is the interpretation Sher intends. Perhaps Scout would *never* give a baby something she thought was dangerous, but received some bad advice about how to calm crying babies which she credulously accepted and now sincerely believes that giving a baby vodka is normal and safe. Sher contends that these agents are blameworthy, despite their unawareness, and I agree.

Given it is so tempting to read improper concern into these agents, I think it is helpful to focus on a more detailed case. Consider the following:

> *Bridge Collapse*: Marie is in charge of conducting a stress test on a local bridge, which involves placing weights on the span to see how it responds. Marie has the option of closing the bridge during the test, which is costly and inconvenient, but sometimes required as a precaution. Whether to close the bridge is a matter of judgment reserved to the engineer; the engineer is charged with determining whether the risk of harm is 'significant' and then determining whether it should be closed based on this classification. Marie is an experienced and responsible engineer, who always closes the bridge when she believes a stress test is too risky, and usually makes the right call.
>
> But as Marie deliberates on this occasion, she commits a serious error of professional judgment. Given the magnitude of the stress test being conducted and the age of the bridge, it would in fact be *highly* imprudent not to close it. Marie ought to know this, given her training and the evidence (which is extremely clear). She should conclude that the risk of collapse is significant and bring to bear her knowledge that in such cases the bridge should be closed. She was fully capable of drawing this inference; but she doesn't.
>
> Tragically, the stress test places too much weight on the bridge span and the bridge collapses, violently killing a pedestrian. Marie blames herself for the death, as does the public, once they realize it was the result of her serious error. It is wholly clear to Marie after

the collapse that her decision to keep the bridge open was highly irrational, a textbook case of poor judgment on the part of an engineer.

Marie is not improperly concerned. For Marie to have acted from improper concern, her failing to close the bridge must have expressed either indifference to moral considerations, or responsiveness to sinister considerations. The relevant moral consideration is the danger that the stress test poses to bridge users. However Marie is not indifferent to the risk she imposes on others; she always closes the bridge when she believes a stress test is too risky. Her pattern of concern is impeccable in this respect. The problem is that she is unaware, on this occasion, that the stress test is too risky, so she doesn't know to close the bridge.

Might her ignorance express improper concern? It would if it were motivated by say, a desire to avoid having to close the bridge on this occasion (perhaps out of laziness). If it were, Marie's ignorance would express a pattern of improper concern consisting of her over-valuing her own convenience, even at the cost of serious risk to others.[49] But Marie's ignorance was simply the result of a concern-neutral cognitive error, not improper concern of this sort.

If one is still tempted to posit improper concern to explain Marie's error, it may be helpful to think of her ignorance as produced by the operation of a cognitive bias (since this can positively replace an explanation that appeals to insufficient concern). Perhaps it is a bright sunny day and Marie has difficulty imagining the possibility of a disaster on such a lovely day. People tend to underestimate the probability of bad outcomes on sunny days because bad outcomes are less likely to come to mind on sunny days, and thanks to the *availability heuristic,* people often

[49] Nor is Marie's ignorance the result of an epistemic vice. FitzPatrick argues that we can be responsible for ignorance that was the result of "voluntary exercise of vices," such as "overconfidence, arrogance, dismissiveness, laziness, dogmaticism, incuriosity, self-indulgence, contempt, and so on" (Fitzpatrick 2008: 605); this is not one of these cases.

estimate the probability of an event based on the ease with which similar events come to mind. [50] Since Marie's ignorance can be fully explained by the operation of this "cold" bias, there is certainly no need to posit improper concern to explain why Marie neither knows nor suspects the stress test is risky.

Perhaps there is another way Marie could be improperly concerned: maybe Marie just doesn't care enough about the rationality of her own beliefs. Rosen observes that "we are under an array of standing obligations to inform ourselves about matters relevant to the moral permissibility of our conduct: to look around, to reflect, to seek advice, and so on" (Rosen 2003: 63). We are not *always* morally obligated to check ourselves for bias; I do nothing morally wrong when I complacently believe that Linda is more likely to be a feminist and a bank teller than a bank teller, or idly question whether the moon landing really took place. But there are cases in which one plausibly should check oneself for irrationality, because the stakes are high, and the cost of checking oneself is low. Maybe Marie displays improper concern by not taking more measures to ensure that her judgment is sound, for instance, by seeking advice.

But it is not plausible that Marie displayed improper concern by not seeking advice or re-checking her belief, because, as stipulated, whether to close the bridge is a matter of judgment reserved to the engineer according to the operative procedural norms. The engineer is given latitude to determine whether the risk of harm is 'significant', so Marie is entirely within protocol when she makes the call based on her own deliberation. It is not plausible that Marie displays insufficient concern for others' safety by failing to go *above and beyond* the protocol, which she has no reason to suspect is flawed. We can even imagine that Marie re-checks her informal calculation, but is subject to the same bias each time, so continues to underestimate the danger posed by the stress test.

[50] It is not that the bias renders Marie incapable of getting the right answer; she is a competent engineer and so fully capable of making the right classification. The sunny day makes the problem slightly harder, but not beyond her power by any means.

There is thus no reason to think that Marie's pattern of concern is defective (and in any cases, we can just stipulate that this is the case). Now the critical question: Are Marie and other comparably irrational agents blameworthy for their foolish decisions? I believe they are. Consider the sort of excuses they might make. Scout in *Colicky Baby* might sincerely plead that she didn't know that vodka was dangerous, that she meant well. Marie might maintain that she followed proper protocol, that she just made an honest mistake. These assertions would all be true, and they justify a certain amount of sympathy for Scout and Marie. But I'm inclined to think that they are only partial excuses, even though they establish the good will of the agents. Marie's foolishness *killed* a person. Scout gave a baby alcohol poisoning. These agents were not *incapable* of exercising better judgment. Marie had never made an incorrect call before, and while a sunny day might affect our judgment, nobody can claim it rationally *debilitated* them and made them incapable of reasoning correctly. And Scout is a normal adult perfectly capable of recognizing that "just give the baby vodka" is poor advice. Our complaint against these agents is simply that they could have and should have known better.

To be clear, my claim is not just that these agents are liable to criticism, or that they should compensate the victims if they can. I think they are fully blameworthy, in the sense of meriting "angry blame": resentment and indignation. To be sure, they merit much less blame than agents who perform comparable acts intentionally. But they are culpable nonetheless.[51]

One might worry that despite every stipulation to the contrary, our sense of culpability is influenced by an inclination to attribute ill will or "mismanagement of opinion" to these agents.

[51] It is also worth noting that Anglo-American criminal law allows liability in some cases of unreasonable ignorance without requiring any further finding of ill will. A successful mistake of fact defense requires the defendant demonstrate his mistake was reasonable: one that an ordinary person would (or might) have made under the circumstances. In Commonwealth v. Pierce (1884), a leading case on ignorance of fact, the defendant was a doctor who attempted to cure a patient by wrapping her in kerosine-soaked clothes. Though the doctor was determined to be acting without knowledge of the risk, he was convicted of involuntary manslaughter on the ground that his ignorance was unreasonable. Such agents may be less culpable than people who cause harm knowingly or recklessly. But it is simply not part of law to excuse them altogether on the ground that they were irrational or stupid or otherwise unreasonable rather than ill willed.

Rosen argues that though we are sometimes inclined to blame agents who act wrongly from blameless ignorance, this is due to a failure to hold clearly in mind all the relevant facts. When we *do* manage to bear in mind that the agent's ignorance was blameless, our resentment evaporates, as Rosen claims in this discussion of our reaction to an ignorant ancient slaveholder (who is unaware that slavery is morally wrong):

> …we are no doubt powerfully inclined to blame [the ancient slaveholder]—so long as we ignore the stipulated fact that he is blameless for not knowing that his slave deserves much more. When we bear this in mind—when we "zoom out", as it were—then (I claim) our sense of his culpability evaporates. It is as if in blaming him we are thinking that he should have known better; he should have known that his action expresses an attitude that would merit intense resentment; he should have known that his act is the sort of act for which one should feel shame in retrospect. When we remind ourselves what it would have taken for him to know these things, and in particular that it's not his fault that he doesn't know them, then our resentment is properly blocked… (Rosen 2003: 73).

So, for instance, if we manage to hold in mind that Marie followed proper protocol—and hence, her ignorance was blameless—our sense of culpability should evaporate according to Rosen. But I do not find Rosen's claim persuasive here. Rosen claims that the "should" thoughts that sustain resentment are falsified when an agent's ignorance is blameless: "he should have known better," "he should have known that his action expresses an attitude that would merit intense resentment," and so on. These thoughts are indeed false if the "should" is interpreted (as it is by Rosen)[52] as indicating the agent should have *done* something differently in the management of her opinion.

[52] According to Rosen, someone is blameworthy her ignorance only if it can be traced to some prior blameworthy act or omission, like a failure to satisfy one's procedural-epistemic obligations. Rosen stipulates that the slaveholder complied with his procedural-epistemic obligations and hence his ignorance is blameless; "we are under no obligation to rethink the uncontroversial normative principles that form the framework for social life" (2003: 65).

But this is not the only way might interpret the relevant "should" thoughts that trigger resentment. If the agent's ignorance is unreasonable, then a different, epistemic version of the "should" thought is available: "X *rationally* should have known better." Her epistemic performance fell below some operative epistemic standard. I maintain that thoughts about the agent's epistemic failings suffice to sustain resentment and indignation in these cases. These emotions are stoked simply by the thought that the agent could have and should have (according to some operative epistemic standard) known better, and this is so even if we expressly cancel the thought that the agent's failure of rationality is fueled by ill will on her part. We imagine the agent having all the evidence in front of her, in full possession of her cognitive faculties, yet drawing an utterly unreasonable conclusion, to dire effect. This is enough to fuel resentment's fire.

Of course, just because we *tend* to resent agents who exercise poor judgment does not mean we are *correct* to resent them. One might dismiss this inclination as seriously misguided, and call for major revision to our ordinary practice of blame and punishment. But I do want to point out that this is in some tension with the Strawsonian approach to moral responsibility that originally motivated the ill will condition. Strawson believed that any credible account of the conditions of responsibility must vindicate ordinary practice at least to a significant degree: our practice of holding one another responsible is an "essential feature of our way of life" and therefore "neither calls for nor permits, an external 'rational' justification" (Strawson 1962: 23). Strawson thought ill will was necessary for blameworthiness because this is what ordinary practice reflects: "The personal reactive attitudes rest on, and reflect, an expectation of, and demand for, the manifestation of a certain degree of goodwill or regard on the part of other human beings towards ourselves" (Strawson 1962: 17). But I've claimed that ordinary practice with the reactive emotions does not always reflect a demand for good will: it sometimes reflects a demand for *reasonableness*, as when we resent agents for their poor judgment. If that's right, then Strawsonians who consider the ill will condition to be vindicated by the standards of ordinary practice have a predicament.

To see why Strawsonians of this sort are in trouble, consider Rosen's Strawsonian view. Rosen is a Strawsonian in the sense that he both (1) endorses the ill will condition and (2) believes this is vindicated by an aspect of ordinary practice, "what the personal reactive attitudes rest on" (Strawson 1962: 347). According to Rosen, resentment does not only *depend* on the belief that the agent expressed ill will, but *constitutively involves* this thought. That is, just like fear of X contains the thought "X is dangerous," resenting X for A contains the thought that "X's doing A expressed an objectionable pattern of concern" (Rosen 2014: 14). And this is why, for Rosen, agents are blameworthy when they express will: expressing ill will makes ordinary resentment appropriate, given its constituent thought. (Just like fear of X is appropriate just in case X is *in fact* dangerous).

But is this right? On a view like Rosen's, we can get a handle on the thoughts implicit in an emotion by asking which sort of discoveries cause it to evaporate. For instance, if fear usually evaporates upon learning or perceiving that X isn't dangerous, this is evidence that "X is dangerous" is a constituent thought of fear. Rosen contends that resentment always evaporates when we "zoom out" and bear in mind the agent's ignorance was blameless. But I've argued that the thought "X could have and should have known better" is enough to sustain resentment, even in the absence of ill will. If that's right, this suggests the ill will thought is not constitutive of resentment.[53] And this is trouble for the foundation of Rosen's account, since it removes the very justification for the ill will condition.

---

[53] One might object to this argument on grounds that it is clearly possible to experience a reactive emotion inappropriately. Fear contains an "X is dangerous" thought, but someone can recoil in fear at the sight of what she knows to be a harmless snake. Might resentment for unreasonableness be like this—possible and understandable but not appropriate?

It is worth noting that even when one knows that a snake is harmless, it can still give the *appearance* of being dangerous—we are hardwired to see the serpentine silhouette as indicating "danger." Hence one might have a quasi-perceptual "X is dangerous" thought, in spite of one's explicit knowledge. But we resent for unreasonableness even when it is *manifestly clear* that the agent was not ill willed (as in *Bridge Collapse*). Though it is always *possible* that our resentment is based on some illusion of ill will, the best explanation for this, I believe, is simply that we can resent without thinking anything like the ill will thought. As Rosen acknowledges: "An account of the thoughts implicit in resentment is unacceptable if it is clear on reflection that we can resent X for A without thinking the thoughts posited by the account" (Rosen 2015: 75).

I concede that these arguments are not decisive. Whether we regard unreasonable ignorance as sufficient for blameworthiness should ultimately depend on the overall plausibility of a theory that rejects the quality of will constraint and provides alternative conditions of blameworthiness. My aim in what follows is to sketch the beginnings of such a theory.

4. Recovering the View of Blameworthiness

Ill will theorists have a pleasingly simple theory of blameworthiness: X is blameworthy for a wrong act A iff in doing A, X expressed ill will (and, perhaps, possessed a certain requisite capacity, like the capacity to assess and respond to moral reasons—what Strawson called a "moral sense") (Strawson 1962). This theory must be revised if we are to account for the culpability of unreasonable but non-ill-willed agents. How might we do so? One possibility is to resign ourselves to an unappealing disjunctive account: X is blameworthy for a wrong act A iff in doing A, X expressed objectionable will, *or* unreasonable ignorance (perhaps, again, against the backdrop of possessing the capacity to assess and respond to reasons). This would be to concede that sometimes we resent agents in virtue of ill will, and sometimes unreasonable ignorance, and that no more unified account of our practice can be given.

But such resignation is unnecessary. There is, I believe, a unified view of moral blameworthiness consistent with our practice which accounts for the blameworthiness of agents who act from unreasonable ignorance, and ill will (though the blameworthiness of ill willed agents will not lie in their ill will, per se, but in their unreasonableness). On this account, agents are blameworthy whenever they act wrongly from an unreasonable attitude—a belief, intention, or desire—that is attributable to them, in a sense I will explain. Acting from unreasonable ignorance and acting from ill will are both ways of acting from unreasonable attitudes attributable to the agent.

To see how cases of blameworthy ill will and blameworthy ignorance might be unified, let us observe that in both sorts of cases, the agent's wrong act reveals an unreasonable attitude

(belief, intention, or desire): one that is held on the basis of bad reasons, or is insufficiently sensitive to the reasons the agent has. This assumes that there are rational norms on final desires as well as beliefs, something that Humeans have long denied. But let us suppose that final desires are rationally criticizable, so that, for instance, a person who finally desires to inflict suffering on others wants something she has no good reason to want. Then wrong action from bad desire and wrong action from irrational ignorance both express unreasonableness, as can be seen in the following pair of cases:

> *Ill Will*: Suppose Smith kills Jones out of malice. There are conclusive reasons for Smith not to kill Jones (just as there are conclusive reasons for anyone not to kill gratuitously), but Smith is irrationally insensitive to these reasons. Hence the *intention* or *desire* that moves Smith to act is unreasonable.

> *Unreasonable Ignorance*: Suppose Smith shoots and kills Jones while horsing around due to a false belief that playing around with a gun isn't risky. Smith has conclusive reason to believe that horsing around with a gun is risky, but is irrationally insensitive to these reasons. Hence the *belief* from which Smith acts is unreasonable.

Of course, the sort of reasons the agent improperly assesses in each case are different. In *Ill Will*, the relevant reasons are reasons for intending or desiring. In *Unreasonable Ignorance*, the relevant reasons are reasons for belief. But insofar as each is subject to rational norms, both agents act from attitudes that are rationally criticizable.

To be sure, in *Ill Will* Smith also expresses ill will in the ordinary sense; he has a final desire to do what is in fact wrong. A traditional Strawsonian will say that this bad desire suffices to make him blameworthy, regardless of whether it is rationally faulty. What I am considering is that it is not the badness of the desire that renders Smith culpable, but its unreasonableness.

Michael Smith (2003) argues for something like this view: it is irrationality in either in an agent's beliefs or desires that renders her "at fault" for wrong conduct. First, consider Gary Watson's contrast between self-indulgence, weakness, and compulsion, which Smith appeals to:

> Suppose that a particular woman intentionally takes a drink. To provide an evaluative context, suppose she ought not to have another because she will then be unfit to fulfill some of her obligations. Preanalytically, most of us would insist on the possibility and significance of the following three descriptions of the case: (1) the reckless or self-indulgent case; (2) the weak case; and (3) the compulsive case. In (1), the woman knows what she is doing but accepts the consequences. Her choice is to get drunk or risk getting drunk. She acts in accordance with her judgment. In (2) the woman knowingly takes the drink contrary to her (conscious) better judgment; the explanation for this lack of self-control is that she is weak-willed. In (3), she knowingly takes the drink contrary to her better judgment, but she is the victim of a compulsive (irresistible) desire to drink. (Watson 1977: 324)

Smith contends that we blame both the self-indulgent woman and the reckless woman. We blame the self-indulgent woman for "having the wrong belief about what she should do in the circumstance" when she could and should have known better. We blame the weak woman, "[not] for her belief — she has the belief she should have, after all—but rather for her failure to act on that belief" when she could have (Smith 2003: 18). By contrast, we do not blame the woman who acted compulsively at all, because she believed correctly and could not have done otherwise than she did.

Having the wrong belief (like the self-indulgent woman) and having the wrong intention or desire (like the weak woman) are both forms of irrationality. The self-indulgent woman fails to recognize the reasons against intending/desiring to take the drink; the weak woman recognizes these reasons, but fails to desire in accord with them. Says Smith: "If people possess, but fail to exercise the relevant [rational] capacities, then there is something that they could have done in a

perfectly mundane sense: they could have exercised the capacity to access the available evidence, or to believe in accordance with their evidence, or to… desire, in the way that's rationally required. Their failure to access or believe or… desire correctly is therefore their fault… and so too is their failure to perform the act that they would otherwise have performed" (2013: 11).

As Smith is careful to emphasize, agents must have the *capacity* to assess reasons, and to modulate their beliefs and desires accordingly, in order to be held responsible for their unreasonable conduct. Nearly everyone agrees that very young children, animals, and the psychotic are not liable to blame even when their acts are sourced in unreasonable intentions, desires, or beliefs. This is presumably because they substantially lack the capacity to assess reasons and to believe and desire in line with assessments of reasons, so their failures to do so are not their fault.

In saying their unreasonableness is not their "fault," we are not yet making a point about moral blameworthiness or the lack thereof in the sense at issue in this paper. As I read him, Smith's talk of 'fault' in this context concerns a different notion: whether the failure is properly attributable to the agent, i.e., whether she is properly criticized qua agent on the basis of it. If an agent possesses the relevant capacity, then — assuming the capacity is not masked (because she has been drugged or is excessively tired or whatnot) — failures to properly exercise the capacity are "down to her." But if she lacks the capacity, they are not "down to her" (they are, rather, attributable to the absence of the capacity), and she is not criticizable for these rational failings.

So, having the relevant capacity—that is, the capacity to assess reasons and modulate one's attitudes in light of these assessments—renders one responsible, *in the attributability sense*, for one's attitudes. That is Smith's point to which I add: agents are blameworthy *in the resentment-involving sense* when they act from unreasonable attitudes that are attributable to them.

Taking account of all this, the resultant theory of blameworthiness, which I will call *Rational Capacitarianism*, is as follows[54]:

> *Conditions of Blameworthiness:* X is blameworthy for a wrong act A iff X's doing A expressed an unreasonable attitude which is attributable to her.[55]

where

> *Theory of Attributability:* An attitude is attributable to X iff X has the capacity to assess the reasons for and against the attitude and modulate her attitude in light of this assessment during the period in which the attitude was formed and sustained.

---

[54] My account bears similarity to Fischer and Ravizza's reasons-responsiveness account of moral responsibility, but should be distinguished. Fischer and Ravizza provide a necessary condition for responsibility: a condition that an agent must satisfy if she is to be morally responsible for her actions. By "responsible," they simply mean an apt target of the reactive attitudes, both negative and positive; "Our Strawsonian view of moral responsibility allows for responsibility for 'morally neutral' behavior. For instance, one can be morally responsible for simply raising one's hand" (1998: 8). In contrast, I'm arguing that *failures* of the capacity for reasons-responsiveness (in the sense I've described) are sufficient for *blameworthiness* when expressed in wrongful action, and moreover, failures of the capacity ground an agent's blameworthiness in cases of wrongful action from *ignorance*. (Fischer and Ravizza don't tailor their account to ignorance; their focus is "the freedom-relevant condition on moral responsibility, rather than the epistemic condition" 1998: 13). Someone could accept Fischer and Ravizza's view that reasons-responsiveness is a necessary condition for responsibility without endorsing my view that failures of this capacity are sufficient for, and moreover are the *ground* of, *blameworthiness* when expressed in wrongful action.

[55] The "expression" of an irrational attitude, we might say, involves its appearance in a certain standard story of action. On the standard story, an agent's φ-ing is an action when it results non-deviantly from her beliefs and desires combining in a certain way, i.e., when she (1) Has some final desire to ψ and (2) A belief that φ-ing is likely to bring about her ψ-ing, which results in an instrumental desire to φ. An action "expresses" an irrational belief or desire when one of the beliefs are desires that produces the action in this way is irrational.

Take again a case of blameworthy ill will: Smith Kills Jones out of malice. On Rational Capacitarianism, Smith is blameworthy because his wrongful act is rooted in some unreasonableness that is down to him:

1. What is the unreasonable attitude expressed? The intention or desire to kill Jones.
2. Why is the intention attributable to Smith? Smith is a competent adult and had the ability to assess reasons for and against intending/desiring, and to modulate his intentions/desires accordingly when the attitude was formed and/or sustained.
3. Why is the intention unreasonable? There are two possibilities. If Smith fails to recognize that there is conclusive reason not to kill Jones despite awareness of the pertinent facts (analogous to the self-indulgent woman), his unreasonableness lies in his flawed belief. If he recognizes there is conclusive reason not to kill Jones but kills him anyway (analogous to the weak woman), his unreasonableness lies in his flawed intention or desire.

Now return to Marie in *Bridge Collapse*. Marie also acts from an unreasonable attitude which is attributable to her: the irrational belief that the stress test isn't risky. Why is the attitude unreasonable? It reflects an improper assessment of the considerations that favor closing the bridge. Why is it attributable to Marie? Marie could have evaluated the evidence properly and believed in accord with it; she was not rationally debilitated. Hence the collapse is Marie's fault, and she is blameworthy for it.

It is not obvious that what makes an attitude attributable to an agent, of course, is that it expresses the failure to exercise a rational capacity of this sort, though this is what bad desires and unreasonable ignorance appear to have in common. One might alternatively maintain that an attitude is attributable to an agent only if it actually reflects her evaluative judgments. This is the view of Angela Smith:

> To say that an agent is morally responsible for something … is to say that that thing reflects her rational judgment in a way that makes it appropriate, in principle, to ask her to defend or justify it … Most of our desires, beliefs, and other attitudes seem to meet this condition of judgment-dependence, even though they do not commonly reflect a choice or decision, and are not normally under our voluntary control (2008: 369-370).

The problem with this view is that it cannot explain responsibility in cases of weakness of will. Both the self-indulgent and the weak woman are responsible in Watson's example, but only the self-indulgent woman's act reflects her evaluative judgment. The weak woman acts *against* her better judgment, taking the drink when she knows she shouldn't.

What unifies the class of attitudes that are attributable to us, I believe, is not that they actually reflect our evaluative judgments, but merely that they are *apt* to reflect them—they are the sorts of attitudes that *could* be brought into alignment with our evaluative judgments. These attitudes, which include beliefs, desires, and intentions, are under a kind of "rational control," in the sense that they are capable of being regulated by, hence controlled by, an agent's recognition of norms, however imperfectly.[56,57] Some writers (e.g. Smith and Pettit) see this sort of "control by the agent's view of the reasons" as strongly analogous to voluntary control, and so as amounting to a kind of freedom akin to free will. But that is not necessary for present purposes. All that

[56] The presence of "rational control" is what leads Pettit and Smith to remark that "[An agent's] beliefs do not just come and go in a natural procession of events…the subject is certainly not a mere passive or mechanical system. She does not just revise her beliefs and desires autonomically, or at any rate, not when they operate beyond the reach of the occasional disabling obstacles that get in her way. She revises them under the spur of recognizing what the relevant norms require of her" (1996: 442)

[57] Some have argued that a kind of capacity-relative control suffices for an agent's ignorance to itself be blameworthy (Rudy-Hiller 2017). For instance, on Rudy-Hiller's Capacitarian account, "an agent can be directly in control (in the responsibility relevant sense) of his ignorance, since he can have direct capacitarian control over whether he notices, remembers, or otherwise is aware of a relevant consideration" (415). And this means an agent can be "directly responsible" for his ignorance (in the sense of being culpable for it) (407). Contra Rudy-Hiller, think agents can be blameworthy for these attitudes themselves; an agent's defective attitude must be expressed in volitional action to be blameworthy.

matters is that when an attitude is formed in an agent who possess the capacity to respond appropriately to reasons and to modulate his attitudes in light of his assessment of the reasons, that attitude is, in every morally relevant sense, his own.

5. Capacitarianism

There is another account of responsibility, which widens the scope of attributable attitudes; on this view, not only failures of rational capacities are attributable to the agent, but failures of any capacity, or perhaps of any "psychological capacity." Capacitarians (Sher 2009, Rudy-Hiller 2017, Clarke 2014, Murray 2018, Amaya & Doris 2015) agree with me that quality of will is not necessary for blameworthiness. Capacitarians contend that agents with a decent quality of will are blameworthy for wrong acts performed from ignorance if, roughly, they could have and should have known the act had the relevant features (Rudy-Hiller 2017: 405)—that is, if their ignorance is the result of the improper exercise of some capacity.

Capacitarians tend to take cases of forgetful negligence as paradigm cases of culpable ignorance, which I have deliberatively avoided doing. For instance, Sher focuses on a case which he calls *Hot Dog*, in which Alessandra forgets her dog Sheba in a hot car. Alessandra forgets about Sheba not because she cares too little about Sheba, but because she gets distracted by a situation occurring at her child's school. Sher contends that Alessandra is blameworthy because she could have and should have remembered that Sheba was in the car; her failure of memory fell below "some applicable" standard which she could have met (Sher 2009: 88).

But we must be careful, as we are not always responsible for failing to exercise capacities we possess. Consider *Cheerleading Disaster*: Sarah is a base on the cheerleading squad tasked with catching Jane, a flyer, after an air flip. Sarah is strong enough to catch Jane and has done so many times before, so she has the capacity to catch her. But on this this occasion, her muscles fail her, and Jane falls through her grip, resulting in serious injury to Sarah. The problem is not a failure intention or effort on Sarah's part; Sarah exerted more than enough effort as would

normally be needed. Her muscles simply don't cooperate with her intention. It is true that Sarah "could have and should have" caught Jane, but this is intuitively insufficient for her moral blameworthiness when we're clear about why she didn't do what she could have and should have done. The challenge for Capacitarians is to distinguish morally relevant capacities from irrelevant ones.[58]

Rational Capacitarian draws the line at rational capacities; all and only failures of rational capacities (including the capacity to form evaluative judgments, and modulate one's attitudes accordingly) are attributable to the agent. The underlying idea is that an agent is responsible only for her attitudes that are apt to reflect her evaluative judgments, that *could* be brought into alignment with them, and would if we were perfectly rational; the attitudes that are under "rational control."

This raises the question: Is what we remember under rational control? Are failures of memory attributable to us; are we answerable for them? This is a difficult question. To be sure, what we remember is sensitive to some degree to our judgments about reasons. Suppose at 8am I think to myself that I really ought to pick up some groceries after work. This makes me more likely to remember to pick up groceries after work. If one judges something to be very important, this increases the chance that thoughts about it will come to mind at the right time, that one's memory will be jogged by things pertaining to it, and so on. If a failure of memory is traceable to a failure to recognize the strength of the reasons for carrying out the task, it is traceable to a failure of a rational capacity and the agent is responsible for it.

The harder case is when an agent recognizes the importance of performing the task, resolves to do it, and then simply forgets to bring about the planned action. This sort of mnemonic

[58] Sher contends that for an agent to be responsible for a failure of a capacity, the failure must be "caused by the interaction of some combination of his constitutive attitudes, dispositions, and traits" (Sher 2003: 88). I'm skeptical that Sher's proposal can explain all the cases. Marie's failure to draw the correct conclusion in *Bridge Collapse* is the result of her being subject to certain biases; and being subject to these doesn't seem constitutive of who she is.

failure seems relevantly different from a failure to intend or believe in accordance with one's reasons. With belief and intention, rationality concerns roughly the synchronic harmony between one's assessments of reasons and one's first order attitudes: ideally rational agents bring them into synchronic consistency. When a rational agent is aware of a reason, forms a judgment about what reason requires, and adjusts her first order attitudes so as to bring them into conformity with this judgment, the episode, though not strictly voluntary, does seem to involve a kind of activity and is properly read as the agent's making the resulting attitudes her own.

Remembering, on the other hand, does not seem to involve a kind of rational activity. Resolving to remember seems more like setting an alarm clock, and hoping it goes off at the right time; the episode of recollection itself either happens or not, though not in a way traceable to the self. Failures of memory seem more like "glitches" analogous to the failure of strength in Cheerleading Disaster than defects in the agent. Consider that in forgetting to perform some task, one's beliefs about the reasons to perform the task are not occurrent, so remembering is not merely a matter of bringing one's assessments of reasons and one's first-order attitudes into synchronic consistency. It seems to me, therefore, that a failure of memory of this sort is not a failure of a rational capacity in the relevant sense, so is not attributable to the agent.

This may be controversial; to the extent that memory is under rational control in the sense in which beliefs, desires, and intentions are, it may make sense to blame agents for failures of memory. But precisely because this is unclear, I think it is a mistake for ill will skeptics to emphasize forgetting cases. When we focus on cases of glitchy memory in good willed agents and stipulate beyond all doubt that at every stage along the way they responded appropriately to the reasons of which they were aware, exercising impeccable rational control, it is hard to see the agent as anything but the passive victim of a glitchy sub-personal module on which he was fully entitled to rely. Cases of this sort are at any rate certainly not clear cases of responsibility. Cases of rational failure on the part of competent agents, I believe, are much clearer. This theory is, of course, only a sketch; but it shows how blame for "ill will" may be more continuous with blame

for unreasonable ignorance than one might have initially supposed. I hope to have shown that we do not need to choose between a quality of will view (which seems too lenient) or a standard Capacitarian view (which seems too harsh). Rational Capacitarianism—on which only rational capacities are relevant—offers a principled alternative.

6. Conclusion

I have argued that agents are blameworthy when they act wrongly from unreasonable ignorance even in the absence of ill will, and provided a theory that ratifies this verdict. The theory entails that when an agent acts wrongly from factual ignorance, she is blameworthy only if her ignorance derives from a failure to exercise a rational capacity—a capacity to recognize reasons for believing and desiring and to be responsive in one's beliefs, desires, or intentions to one's assessment of reasons.

The plausibility of a theory that licenses blame for non-ill-willed unreasonableness in part depends on what blame and blameworthiness consists in. In closing I will say a few words about this issue. Like Rosen (2015), Gibbard (1992), and Wallace (1994), I have fastened on a conception of blame according to which the primary forms of blame are the negative reactive emotions like resentment and indignation, and on which an agent is blameworthy for an act iff she is liable to resentment and indignation for it. But this isn't a complete analysis, because it doesn't specify what being "liable" to the negative reactive attitudes amounts to. How we fill this gap is important.

Rosen's view, for instance, is that since resentment and indignation are emotions, their appropriateness conditions are given by their constitutive thoughts. Emotions involve belief-like mental states; fear involves the thought that X is dangerous, misery the thought that things are going badly, and so on, and an emotion is appropriate just in case its constituent thoughts are true (Rosen 2015: 71-72). Accordingly, for it to be appropriate to resent X for A is for the thoughts implicit in resentment to be true of X and A (Rosen 2015: 72).

Everything I have said in this paper is consistent with this general view. If we adopt this framework, I should be read as contesting Rosen's view of the constitutive thoughts of resentment. Rosen thinks there is a "quality of will thought" contained in resentment ("in doing A, X expressed an objectionable pattern of concern"), and I disagree. My argument for this is simply that it is possible to resent X for an act that expresses unreasonable ignorance, but not ill will, clearheadedly, in full awareness of this fact. Resentment is stoked in these cases by the thought that X could have and (rationally) should have known better.[59]

If we reject Rosen's account of what it is to be "liable" to the reactive emotions, it may be harder to see how resentment directed at unreasonableness could be appropriate in the absence of ill will. The "fairness view" (Wallace 1994) specifies that X is liable to resentment for A just in case it would be morally fair to resent her for A, and the "fittingness view" appeals to a primitive, *sui generis* relation of fittingness (see Rosen 2015: 70). But how could it be fair, or fitting, to resent—which involves the withdrawal of some amount of good will—if there was no comparable ill will on the part of the target of the resentment?

I do not have a complete answer to this question, but I will outline the beginnings of a response. The fairness of resentment plausibly depends on whether the target of the resentment had a fair opportunity to avoid being a target of resentment. Rudy-Hiller contends, for instance, that blaming for ignorance is appropriate because an agent "can have direct capacitarian control over whether he notices, remembers, or otherwise is aware of a relevant consideration" (2017: 415), which is simply a matter of having the "requisite abilities to do these things and a fair opportunity to exercise them" (407). I differ from Rudy-Hiller in thinking that resentment is only appropriately directed at attitudes which are under the agent's *rational* control, but the spirit of my view is the same. Agents can avoid being the target of resentment by exercising their capacity

[59] As Rosen acknowledges, "An account of the thoughts implicit in resentment is unacceptable if it is clear on reflection that we can resent X for A without thinking the thoughts posited by the account" (Rosen 2015: 75).

to recognize reasons for believing and desiring and to be responsive in their beliefs and desires to their assessment of reasons. This is enough, it seems, for a fair opportunity to avoid.

Resentment plausibly is a sort of sanction directed at agents who we think could have done better, aimed at ensuring the agent will do better in the future. The attitudes with respect to which this corrective activity makes sense are precisely the ones that are under the agent's rational control—the attitudes which she can modulate according to her evaluative judgments. What resentment "wants," so to speak, is for the agent to recognize and revise her defective attitude (for instance, by adopting the appropriate belief or desire); this is why resentment is often not satisfied unless the agent recognizes *why* what she did was wrong. If resentment is a correctional activity of this sort, it is easier to see how it could be appropriately directed at all forms of unreasonableness, not just ill will.

I will conclude by stressing that the justification of serious punishment is another matter. Resentment and indignation are moderate, temporary responses which tend to lessen when the agent corrects the attitudes toward which resentment is directed. Serious punishment (e.g., a restriction of agent's liberty) of the sort that conveys serious moral condemnation is plausibly inappropriate unless the agent expressed improper concern; as the costliness to the agent of a blaming response increases, the more it seems reasonable to require some morally objectionable feature of the agent's will to be involved.[60]

[60] Thanks to an anonymous reviewer for this point.

References

1) Amaya S, Doris JM. No excuses: Performance mistakes in morality. 2015.

(2) Arpaly N. Unprincipled virtue: An inquiry into moral agency. : Oxford University Press; 2002.

(3) Baker D, Woods J. How expressivists can and should explain inconsistency. Ethics 2015;125(2):391-424.

(4) Beddor B. Subjective disagreement. Noûs 2018.

(5) Bertea S. Why One's Practical Reasons are not just One's Own Private Affair. Philosophical Inquiry 2017;41(1):63-85.

(6) Bratman M. Intention, plans, and practical reason. : Harvard University Press Cambridge, MA; 1987.

(7) Bratman ME. Shared agency: A planning theory of acting together. : Oxford University Press; 2013.

(8) Bratman ME. Thinking How to Live and the Restriction Problem. Philosophy and Phenomenological Research 2006;72(3):707-713.

(9) Bratman ME. Valuing and the Will. Philosophical Perspectives 2000;14:249-265.

(10) Bratman ME. Shared cooperative activity. The philosophical review 1992;101(2):327-341.

(11) Clarke R. Negligent action and unwitting omission. Surrounding free will 2014:298-317.

(12) Cullity G. Thinking How to Live–Allan Gibbard 2007.

(13) Darwall SL. Impartial reason. 1983.

(14) Dreier J. Accepting agent centred norms: A problem for non-cognitivists and a suggestion for solving it. Australasian Journal of Philosophy 1996;74(3):409-422.

(15) Egan A. There's something funny about comedy: A case study in faultless disagreement. Erkenntnis 2014;79(1):73-100.

(16) Egan A. De gustibus non disputandum est (at least, not always). Disagreement.Oxford University Press (forthcoming) 2010.

(17) Finlay S. Confusion of tongues: A theory of normative language. : Oxford University Press; 2014.

(18) Finlay S. Normativity, necessity, and tense. A recipe for homebaked normativity. Oxford studies in metaethics 2010;5:57-85.

(19) Finlay S. What ought probably means, and why you can't detach it. Synthese 2010;177(1):67-89.

(20) Fischer JM, Ravizza M. Responsibility and control: A theory of moral responsibility. : Cambridge university press; 1998.

(21) FitzPatrick WJ. Moral responsibility and normative ignorance: Answering a new skeptical challenge. Ethics 2008;118(4):589-613.

(22) Gert J. Korsgaard's private-reasons argument. Philosophy and Phenomenological Research 2002;64(2):303-324.

(23) Gibbard A. Reply to critics 2006.

(24) Gibbard A. Normative and recognitional concepts. Philosophy and Phenomenological Research 2002;64(1):151-167.

(25) Gibbard A. Wise choices, apt feelings: A theory of normative judgment. : Harvard University Press; 1990.

(26) Gibbard A. A noncognitivistic analysis of rationality in action. Social Theory and Practice 1983;9(2/3):199-221.

(27) Gibbard A, Gibbard A. Thinking how to live. : Harvard University Press; 2009.

(28) Gilbert M. Walking together: A paradigmatic social phenomenon. MidWest studies in philosophy 1990;15:1-14.

(29) Graham P. Blame, Determinism, and Ignorance 2006.

(30) Gregory A. Are All Normative Judgments Desire-like. J.Ethics & Soc.Phil. 2017;12:29.

(31) Hare RM. Meaning and speech acts. The Philosophical Review 1970;79(1):3-24.

(32) Harman E. Does moral ignorance exculpate? Ratio 2011;24(4):443-468.

(33) Kahneman D, Slovic SP, Slovic P, Tversky A. Judgment under uncertainty: Heuristics and biases. : Cambridge university press; 1982.

(34) Kalin J. In defense of egoism. Morality and Rational Self-Interest 1970:64-87.

(35) Korsgaard C. The reasons we can share: An attack on the distinction between agent-relative and agent-neutral values. Social Philosophy and Policy 1993.

(36) Korsgaard CM, Korsgaard CM, Korsgaard CMK, Cohen GA, Geuss R, Nagel T, et al. The sources of normativity. : Cambridge University Press; 1996.

(37) LeBar M. Korsgaard, Wittgenstein, and the mafioso. The Southern journal of philosophy 2001;39(2):261-271.

(38) Lewis D. Attitudes de dicto and *de se*. The philosophical review 1979;88(4):513-543.

(39) Manne K. Internalism about reasons: sad but true? Philosophical Studies 2014;167(1):89-117.

(40) McKenna MS. The limits of evil and the role of moral address: A defense of Strawsonian compatibilism. The Journal of Ethics 1998;2(2):123-142.

(41) Medlin B. Ultimate principles and ethical egoism. Australasian Journal of Philosophy 1957;35(2):111-118.

(42) Murray S. Responsibility and vigilance. Philosophical Studies 2017;174(2):507-527.

(43) Nagel T. The possibility of altruism. : Princeton University Press; 2016.

(44) Pettit P, Smith M. Freedom in belief and desire. The Journal of Philosophy 1996;93(9):429-449.

(45) Richard M. Relativistic content and disagreement. Philosophical Studies 2011;156(3):421-431.

(46) Ridge M. Expressivism and Collectives. Mind 2018;127(507):833-861.

(47) Ridge M. Reasons for action: agent-neutral vs. agent-relative. 2005.

(48) Ridge M. Disagreement. Philosophy and Phenomenological Research 2013;86(1):41-63.

(49) Rosen G. The alethic conception of moral responsibility. The nature of moral responsibility: New essays 2015;1:65-88.

(50) Rosen G. Skepticism about moral responsibility. Philosophical Perspectives 2004;18(1):295-313.

(51) IV—Culpability and Ignorance. Proceedings of the Aristotelian Society (Hardback): Wiley Online Library; 2003.

(52) Ross J, Schroeder M. Reversibility or disagreement. Mind 2013;122(485):43-84.

(53) Ross J, Schroeder M. Reversibility or disagreement. Mind 2013;122(485):43-84.

(54) Roth AS. Intention, expectation, and promissory obligation. Ethics 2016;127(1):88-115.

(55) Roth AS. Prediction, authority, and entitlement in shared activity. Noûs 2014;48(4):626-652.

(56) Roth AS. Shared agency and contralateral commitments. The Philosophical Review 2004;113(3):359-410.

(57) Rudy-Hiller F. A capacitarian account of culpable ignorance. Pacific Philosophical Quarterly 2017;98:398-426.

(58) Scanlon T. Forms and conditions of responsibility. The nature of moral responsibility: New essays 2015:89-114.

(59) Scanlon T. Reasons and decisions. Philosophy and Phenomenological Research 2006;72(3):722-728.

(60) Schroeder M. Ought, agents, and actions. Philosophical Review 2011;120(1):1-41.

(61) Schroeder M. Being for: Evaluating the semantic program of expressivism. : Oxford University Press; 2008.

(62) Schroeder M. Slaves of the Passions. : Oxford University Press; 2007.

(63) Searle J. Collective intentions and actions, intentions in communication. PR Cohen, J.Morgan and ME Pollack (Eds) 1990:401-416.

(64) Shafer-Landau R. Oxford studies in metaethics. : Oxford University Press; 2006.

(65) Sher G. Who knew?: Responsibility without awareness. : Oxford University Press; 2009.

(66) Sidgwick H. The methods of ethics. : Good Press; 2019.

(67) Smith AM. Responsibility for attitudes: Activity and passivity in mental life. Ethics 2005;115(2):236-271.

(68) Smith M. A constitutivist theory of reasons: Its promise and parts. Law, Ethics and Philosophy 2013:9-30.

(69) Soames S. Presupposition. Handbook of philosophical logic: Springer; 1989. p. 553-616.

(70) Strawson PF. Freedom and resentment and other essays. : Routledge; 2008.

(71) Wallace RJ. The publicity of reasons. Philosophical Perspectives 2009;23:471-497.

(72) Wallace RJ. Responsibility and the moral sentiments. : Harvard University Press; 1994.

(73) Watson G. Skepticism about weakness of will. The Philosophical Review 1977;86(3):316-339.

(74) Worsnip A. 'Ought'-contextualism beyond the parochial. Philosophical Studies 2019;176(11):3099-3119.

(75) Zimmerman MJ. Moral responsibility and ignorance. Ethics 1997;107(3):410