Northwestern University Society for the Theory of Ethics and Politics

13th Annual Conference

May 9-11, 2019
John Evans Alumni Center
1800 Sheridan Road
Evanston, IL

This conference is supported by the generosity of:
The Philosophy Department
The Weinberg College of Arts and Sciences
The Kreeger Wolf Endowment
The Graduate School
The Alice Kaplan Institute for the Humanities
The Brady Scholars Program in Ethics and Civic Life

Table of Contents

SCHEDULE…….…..... 2
Acknowledgements…….....4

"Hypocrisy, impartiality, and standing"…….5
Max Kramer (Arizona)

"Agency, Akrasia and the Normative Environment"……12
Gregory Antill

"The Importance of Logically Complex Actions" ……..21
Andrew Flynn

"Personal and Impersonal Good"…….39
Nandi Theunissen

"Moral Liability in War and Self-Defense: Extending the Just Cause Argument"…….48
Katherine Sweet

"You're So Smug, I'll Bet You Don't Care this Paper is About You"……54
Grant Rozeboom

"Addiction as Evidence"……..66
Cami Koepke

Keynote Address:
"A Convenient Truth? Subjective Well-Being and Global Climate Change"…….73
Peter Railton

"On what is "out of the question"……..117
Lilian O'Brien

"Self-Mastery in Plato's Laws" ……...……..127
Brian Reese

"Moral Responsibility and Cooperation"……..……143
Saba Bazargan-Forward

Keynote Address:
"What, if Anything, is Disagreement in Attitude?"……..151
Sarah Stroud

Speaker Bios …….…….…..…… 166
Chicago Attractions…….168
Map of Downtown Evanston…….....….... 169

Northwestern University Society for the Theory of Ethics and Politics
13th Annual Conference
May 9-11, 2018
John Evans Alumni Center

Thursday, May 9th

9:00-10:25
"Hypocrisy, impartiality, and standing"
Max Kramer (Arizona)
Comments by Hao Liang (Northwestern)

10:35-12:00
"Agency, Akrasia and the Normative Environment"
Gregory Antill (Claremont McKenna)
Comments by Amy Flowerree (Texas Tech)

Lunch

2:15-3:40.
"How I Spent My Summer Defending-or-Defeating Anscombe: Anscombian Action Theory and the Possibility of Logically Complex Actions"
Andrew Flynn (UCLA)
Comments by John Beverley (Northwestern)

3:50-5:15
"Personal and Impersonal Good"
Nandi Theunissen (Pittsburg)
Comments by Gwen Bradford (Rice)

Dinner

Friday, May 10th

9:00-10:25
"Moral Liability in War and Self-Defense: Extending the Just Cause Argument"
Katherine Sweet (SLU)
Comments by Michael Schwarz (Northwestern)

10:35-12:00
"You're So Smug, I'll Bet You Don't Care this Paper is About You"
Grant Rozeboom (St. Norbert)
Comments by John Lawless (Davidson)

Lunch

2:15-3:40
"Addiction as Evidence: Frankfurt's Unwilling Addict and the Explanatory Gaps of Mesh Theories of Responsibility"
Cami Koepke (UCSD)
Comments by Christiana Eltiste (Northwestern)

3:50-5:45
Keynote Address:
"A Convenient Truth? Subjective Well-Being and Global Climate Change"
Peter Railton (Michigan)
Comments by Wendy Salkin (San Francisco State)

Reception – Everyone is invited

Saturday, May 11th

9:00-10:25
"On what is "out of the question"
Lilian O'Brien (Helsinki)
Comments by Kristina Gehrman (Tennessee)

10:35-12:00
"Self-Mastery in Plato's Laws."
Brian Reese (Penn)
Comments by Andy Hull (Northwestern)

Lunch

2:15-3:40
"Moral Responsibility and Cooperation"
Saba Bazargan-Forward (UCSD)
Comments by Allen Coates (ETSU)

3:50-5:45.
Keynote Address:
"What, if Anything, is Disagreement in Attitude?"
Sarah Stroud (UNC)
Comments by Mike Zhao (NYU)

Dinner

Acknowledgements

Conference Organizers:

Kyla Ebels-Duggan, Richard Kraut, Stephen White, and Hao Liang

For organizational and administrative assistance, special thanks to Crystal Foster, Thomas Winters, Casey Huynh and the Graduate Students in the Department of Philosophy

NUSTEP is generously supported by:
Alice Kaplan Institute for the Humanities
The Brady Scholars Program
The Department of Philosophy
The Graduate School of Northwestern University
The Edith Kreeger Wolf Endowment
Weinberg College of Arts and Sciences

Hypocrisy, impartiality, and standing

Max F. Kramer
University of Arizona
maxkramer@email.arizona.edu

Much attention has recently been paid to hypocrisy in the literature on moral responsibility, with good reason. Especially in the United States, we are in a cultural moment where charges of hypocrisy are leveled at nearly all of our politicians, not only by onlookers but by other politicians as well. As such, we want to be sure that those charges are directed at the right people, in response to the genuine phenomenon, with the right effects following from the charge. Here I defend the non-hypocrisy (NH) condition on standing to blame and vindicate the basic structure of Macalester Bell's taxonomy of hypocrites in the face of challenges from Kyle Fritz and Daniel Miller.

The standing to blame

Let's start with a basic observation: our license to participate in activities is based on a variety of factors. If I found myself, in what would be a concerning turn of events, not anesthetized in an operating room, I could not just mosey over to the table and try my hand at performing a surgery. In one sense I *can*, of course—I do not lack the physical ability to walk or to hold surgical tools or even to cut into human flesh. We can also stipulate that I am not being stopped by any doctors (imagine them awestruck), nor am I internally constrained by faintness as the sight of blood or hesitation to cut into a body. What says that I cannot perform surgery? When we are answering this question, we are explaining why I do not have *standing* to perform surgery. The standing to participate in some activity is akin to a license or a credential. When you meet certain criteria, you can participate; if not, it is impermissible for you to do so. Importantly, this 'licensing' conception of standing need not be elitist. It is hard to find a living thing around that lacks the standing to breathe or to drink water. But, the license metaphor is useful because it illustrates that standing is something that can be revoked.

There are two general ways to lose standing in a domain. Consider a game of pickup soccer. Anyone in the park can join—merely being around and physically able to participate is enough to get on the pitch. One way to lose this standing is to voluntarily give it up. If I quit and tell people that I'm going home, I can no longer participate in the game. I call it *forfeiting* standing when standing is voluntarily relinquished by an agent. Once a team says they've forfeited, they can no longer keep trying to win the game—they gave up their standing to do so of their own volition. However, sometimes an agent loses standing involuntarily. Soccer, like all activities, is governed by a set of rules, some of which are constitutive of the activity.[1] If I violate the constitutive rules by which the game is played, then I must be held accountable for that malfeasance and stripped of my standing. I call it *undermining* standing when standing is stripped from an agent for their failure to adhere to the principles that constitute some activity. An agent can undermine their standing (and equivalently, fail to adhere to governing principles) in two ways. One way is to demonstrate a failure to understand the rules. The soccer game in the park had very low entry criteria. However, when I get on the pitch and immediately pick the ball up with my hands, run over to the basketball courts, and start shooting three-pointers, I lose my standing to play in the soccer game. This is because I've made the game impossible for the other players to play, but notice that this is so because I've shown myself to be ignorant of certain constitutive rules in the absence of which we'd no longer be playing something recognizable as soccer. It's impossible to play when even one participant doesn't understand what's going on. Perhaps in

[1] These constitutive rules should be understood in the same way as Rawls' 'practice conception' of rules (Rawls 1955, 24).

the future I can demonstrate that I've learned the rules, but for now, I have undermined my standing. Call such a case *undermining by ignorance*. On the other hand, I may know and understand the rules of soccer, but refuse to play by them because it gives me an advantage. In soccer, one cannot touch the ball with one's hands, but it sure does make it a lot easier to score. If I am caught doing this, my teammates and opponents alike will kick me off the field, and they will have needed to, because again, the game cannot be played unless everyone follows the rules. Call such a case *undermining by cheating*.

The standing to blame, which is the standing thought to be lost by engaging in hypocrisy, works in the same way. Being a hypocrite is not a signal of forfeit, so when we talk about NH, we are talking about undermining one's standing to blame as a result of engaging in hypocritical behavior. The thought behind NH is that to be a hypocrite is to violate some constitutive rule of morality. There are two ways to come by this violation—one either doesn't understand the principles of morality or one recognizes them but contravenes them to get an upper hand. The hypocritical agent is either ignorant or a cheater, and either strip them of their standing to blame others.

Hypocrisy

Based on what I have said in the previous section, hypocrisy undermines an agent's standing to blame in the case that it reveals her to either be ignorant of the constitutive rules of morality or purposefully in violation of them. To determine whether this is the case, we need a characterization of hypocrisy. Macalester Bell originally defines NH in terms of the agent "not having engaged in similar wrongdoing in the past."[2] As a variety of authors have pointed out, this characterization seems insufficient. It does not seem hypocritical to, for instance, have lied in the past and then to blame someone else for lying if one has appropriately made up for one's past transgression. Our characterization must be more precise. Hypocrisy generally involves employing a double standard in applying moral norms. This is what Bell seems to be getting at by making reference to an agent's previous wrongdoing. The hypocrite says that some violation is wrong for others (perhaps specific others), but not wrong for her (or for some specific others, such as a group to which she belongs).[3] More precisely, there is an inconsistency in the agent's blaming practices, including her disposition to blame others, to blame herself, and to accept blame from others. This marker of inconsistency is frequently noted in the literature.[4] R. Jay Wallace is perhaps the most explicit about this: "A charge of hypocrisy […] purports to isolate an internal inconsistency of some kind in your views."[5] However, neither Wallace nor Bell think that inconsistency alone is enough to supply the proper kind of weight that we attribute to hypocrisy as a moral vice.[6]

[2] Bell (2012), p. 264.

[3] Hypocrisy is often talked about in the responsibility literature as a difference in norm application between the agent and others, but when ordinary people talk about the hypocrisy of politicians, for example, they make reference to a phenomenon of turning a blind eye to a party member's malfeasance while raising a stink about the same violation by a member of an opposing party. This leads me to believe that hypocrisy is less narrowly constrained to what an agent will take responsibility for, but for simplicity's sake, I will ignore the possibility of hypocrisy on behalf of a third party.

[4] Wallace (2010), Fritz and Miller (2018), Rossi (2018), Roadevin (2018).

[5] Wallace (2010), p. 307.

[6] A note of Wallace's is of interest here, for it foreshadows what I will say below. "Kantians, of course, characterize moral requirements as ones that we have to comply with on pain of a kind of inconsistency. But this aspect of the Kantian position is an attempt to explicate the rational or normative importance of moral requirements, not an account of the moral objection that can be brought against behavior that violates those requirements" (Wallace 2010, p. 310 n. 6).

Since we are struggling to find a formal definition of hypocrisy, it may be useful to look at some different types of hypocrites. What all these hypocrites have in common, presumably, will constitute the core of hypocrisy. Bell[7] provides a useful taxonomy consisting of three types of hypocrites: the clear-eyed, the exception-seeking, and the weak-willed. When we take issue with cynical politicians who act in bad faith and display crocodile tears to gin up their voter base, we are responding to their clear-eyed hypocrisy. Clear-eyed hypocrites "only pretend to care about the norms in question and feign the negative affective attitudes at the heart of blame."[8] Clear-eyed hypocrites know that they are being inconsistent, but do it anyway. By contrast, the exception-seeker "really does care about moral standards and genuinely harbors hard feelings for those who do wrong. Nevertheless, he engages in the same kind of behavior that he criticizes and sees his own behavior as morally justified."[9] The exception-seeking hypocrite is wrongfully ignorant of his inconsistency. Finally, the weak-willed hypocrite acknowledges that she is wrong in violating certain moral norms—she feels guilt and remorse—but can't, due to her weakness of will, keep herself from violating them. She also, despite her own issues, persists in blaming others for their violations of the same norms.

All three of these characters strike me as genuine hypocrites and it seems plausible to me that their hypocrisy is grounds for them being stripped of their standing to blame others for the moral norm violations at issue in their hypocrisy. However, recent defenses of NH identify the clear-eyed and weak-willed as counterfeit hypocrites, at least under a construal of hypocrisy needed for NH to be defensible. This is not entirely surprising from a dialectical standpoint—Bell rejects NH, largely because she thinks that the moral wrongs at the root of hypocrisy are insufficient to undermine an agent's standing to blame, and that blame has multiple aims, some of which are not influenced by an agent's bad (even vicious) character.

Blame and the Hostile Attitudes account

Thus far, I have said nothing substantive about the nature of blame. A popular account, and one to which many of the authors in this debate have hitched their boat, is the Hostile Attitudes account of blame (HA).[10] HA says that, some agent *R* blames agent *S* for action *A* only if:

(1) *R* believes that *S* is an agent of *A*.

(2) *R* believes that *S*'s *A*-ing is wrong or bad.

(3) *R* believes that *S* is blameworthy for *A*-ing.

(4) *R* experiences negative emotions (indignation, resentment, contempt, guilt) on account of (1), (2), and (3).[11]

[7] Bell (2012), p. 275-276.
[8] *Ibid.*, p. 275.
[9] *Ibid.*, p. 276.
[10] Bell (2012), Fritz & Miller (2018), Wallace (2010), and Rossi (2018), for example, all subscribe to some version of this account.
[11] This is the formulation given by Rossi (2018, p. 554). Bell (2012) gives the account as a conjunction of (3) and (4), presumably because one would think that (1) and (2) are entailed by (3).

HA, then, is a thesis that negative reactive attitudes are *constitutive* of blame—no occurrent emotion, no blame. This immediately casts doubt on the viability Bell's taxonomy; clear-eyed hypocrites do not blame others at all. If one does not care about the moral norm in question, then it is unlikely that one would feel indignation or contempt for another whom one is accusing of violating that norm. Since on HA, the reactive attitude is constitutive of blame, HA says that the clear-eyed do not hypocritically *blame*. At best, they do something else, like morally grandstand.[12] This is not immediately a problem—there remains an inference to be made between a purported hypocrite not engaging in blame to the agent not being a hypocrite, but it is an inference that seems eminently plausible to make.

Kyle Fritz & Daniel Miller[13] make just such an inference, arguing that since the clear-eyed 'hypocrite' does not genuinely blame, then she is not a hypocrite at all. They argue that the defining feature of a hypocrite is that they have a *differential blaming disposition* (DBD). This means that the hypocrite is disposed to blame others for violations of some moral norm, but not disposed to blame herself for the same violations, with no justifiable reason for this difference.[14] The DBD account disqualifies both the clear-eyed and the weak-willed from hypocrisy: clear-eyed 'hypocrites' do not blame others at all (on HA), so they do not differ in their blaming dispositions when it comes to themselves and others, and weak-willed 'hypocrites' do blame themselves, so they also do not show a difference in disposition.

Partiality at the root of hypocrisy

Fritz & Miller (along with Wallace) take the wrongdoing of the hypocrite to be a denial of equal moral worth between agents, which is a fundamental principle of morality. This denial of equal worth is implicit in an agent's having a DBD. Here is their argument in brief. (1) Hypocrisy is equivalent to having a DBD; (2) to have a DBD rejects the impartiality of morality; (3) rejecting impartiality is equivalent to rejecting equality of persons; (4) if one rejects equality of persons, one rejects what gives one the right to blame others; (5) if one rejects the grounds of blame, one forfeits the right to blame others; so, (6) if one is a hypocrite, one forfeits the right to blame others.[15]

The first point to make is that Fritz & Miller use the term 'forfeit' in a different way than I have above in my discussion of standing. I use 'forfeit' to mean something like a voluntary giving up (of one's standing in a domain), while Fritz & Miller talk about forfeiting the right to blame. Forfeiture in their sense is something that is imposed on the agent, while in my sense, it is something that an agent voluntarily undertakes. I prefer my own usage (shockingly!) because I prefer to keep the distinction between rights or standing that an agent voluntarily gives up and those that are taken away from the agent. However, it should be quite clear that Fritz & Miller are talking about the same phenomenon I am talking about when I say that an agent's standing to blame is undermined by her actions.

More importantly, the argument is valid, and more importantly still, it expresses what I take to be the core of their (and Wallace's) analysis of hypocrisy: hypocrisy is wrong because morality is fundamentally impartial and hypocrisy constitutes an immoral way of being partial toward oneself. The sort of inconsistency going on, which Fritz and Miller characterize as a differential blaming disposition, is an inconsistency in the way one treats oneself compared to others. Because morality is impartial, this kind of inconsistency is illicit, and engaging in it undermines an agent's standing. Treating oneself and others impartially when it comes to blame or any other aspect of morality is fundamental because in the moral

[12] Tosi & Warmke (2016).
[13] Fritz and Miller (2018).
[14] Note that Fritz and Miller choose to focus on the agent's disposition to blame herself, not her disposition to accept blame.
[15] *Ibid*., p. 125.

realm, we are all on equal footing—all persons are equally subject to moral norms. To connect back to my analysis of standing, hypocrites on this view undermine their standing precisely in the way I suggested: they in some way fail to respect the constitutive of the activity in which they participate, viz., blame or moral address more generally. Moreover, this is an analysis with which Bell ought to agree. She views hypocrisy as an attitude of superbia; that is, one that involves holding oneself above others, morally speaking. If equality is a constitutive rule of morality, then hypocrisy, as a species of superbia, should undermine standing.

Impartiality and universality

I said at the outset that my aim was to defend NH and to vindicate Bell's taxonomy of hypocrites. So far, though, all I have done is provided some evidence that Bell's taxonomy is faulty and that Fritz and Miller's defense of NH fits into the more robust account of standing I provided at the outset. What is left is to show how Fritz and Miller's account is lacking, and that the way to fix it also rules back in the hypocrites that HA and DBD ruled out.

A trivial way of doing this is to reject HA. If HA is not in play, and blame is instead defined functionally, then the clear-eyed could be ruled back in as hypocrites, because (on one possible view) they call other agents to account for some norm violation, which is the function of blame. I think that this is the wrong way to go for two reasons. One, I take it there are independent reasons to endorse HA, so I need not alienate most other interlocutors by endorsing some competing view. And two, this move does not rule back in weak-willed hypocrites, and therefore I would have failed in my vindicatory project. Instead, what I plan to do is question Fritz and Miller's understanding of impartiality and then show that what is illicitly partial about hypocrisy is present in all three of Bell's hypocrites, though the akratic hypocrite needs further characterization.

Impartiality plays a role in (2) and (3) of Fritz and Miller's argument. They say that partiality violates the equality of persons that is essential to morality. However, this needs further explication, because it is quite plausible that some forms of partiality are part and parcel of morality, and do not violate the equality of persons. For example, it is partial, but likely permissibly so, to privilege saving one's child over a stranger's child in an emergency situation (provided you can only save one). It is also permissibly partial to show friends more trust than you would give strangers, and not just for epistemic reasons. These forms of partiality are only problematic on certain accounts of morality. Consider Mill, who was adamant that agents must be strictly impartial with regards to the interests of each person involved in a utilitarian calculus. Mill would strike preferential treatment down as indeed immoral. This seems to be the sense of impartiality operative in Fritz and Miller's minds—equality of persons means that each person's interests count equally when it comes to blame. This is why it seems that the clear-eyed and the weak-willed are not hypocritical: in strange ways (or at least, not in hypocritical ways), they actually do not treat their interests in avoiding blame as different than others'.

However, there is a competing way to understand impartiality, which is the idea that for an action to be permissible, it must be permissible for all agents. This is the sense in which Kant's Formula of Universal Law is impartial. Kant thinks that, unless your maxim can be valid for all rational agents simultaneously without resulting in contradiction, it is impermissible to act on it.[16] Kantian impartiality, better termed universality, plausibly permits those forms of kin- and friend-favoritism, because it permits them as long as all moral agents can avail themselves of the same favoritism without morality collapsing.

[16] There are a variety of competing interpretations of the Universal Law Formula, but I expect that this conveys the gist well enough.

This is precisely the test of constitutive rulehood—once we drop universality, we can't have equality, and we've already agreed that without equality, we're no longer talking about morality. The same is not true of impartiality. Therefore, it looks as though Kantian universality, not Millian impartiality, is what is needed to guarantee the equality of persons required by morality.

When we return to Fritz & Miller's analysis with this result in hand, we find that their account is myopic in excluding the clear-eyed and the weak-willed. Although the clear-eyed and the weak-willed do not differ in their dispositions to blame themselves and others, they *do* require others to accept blame as genuine (and all that entails) that they do not themselves accept as genuine. For the victim of the clear-eyed hypocrite, the blame they receive is seen as genuine, because it calls them to account for their violation and take steps to remedy the consequences they generate. Meanwhile, the clear-eyed hypocrite will refuse accountability for the same violation, and this is the failure to play by the rules of morality that undermines their standing to blame, not whether or not their blame is genuine. By turning our view from blaming to accepting blame as genuine, we see that clear-eyed hypocrites do genuinely hold themselves to different standards than they hold others in a way that violates the constitutive rules of morality, which is what is at issue in whether hypocrisy undermines the standing to blame.

How fares the person of weak will? It seems as though she does accept blame for her actions just as she expects others to, and is therefore not improperly partial. But I think the case requires more characterization. Compare the following two figures:

> **Alice** is a serial cheater. She genuinely tries to stay faithful to her partner, but her urge for sex always wins out while her partner is away on long business trips. This causes serious problems in her relationship, and she has frequently lamented to you that she wishes her willpower were stronger. In conversation with Alice, you let slip that you recently cheated on your own partner. Alice admonishes you and tells you that what you did was wrong.

> **Beatrice** is a serial cheater. Because of her desire for sex, she sleeps with other people while her partner is away on long business trips. She understands that these liaisons are morally wrong and feels guilty about them, but resigns herself to the idea that her sex drive is too strong, and her will is too weak, to try to change her behavior. In conversation with Beatrice, you let slip that you recently cheated on your own partner. Beatrice admonishes you and tells you that what you did was wrong.

To report my own intuitions, I find Beatrice's criticism much more objectionable than Alice's. I am disposed to take Alice's criticism to heart, while I find Beatrice to be criticizing me without standing. There is nothing in the foregoing account of weak-willed hypocrisy to suggest that we should feel differently about these two cases. Both of them blame you for a violation of a norm that they frequently violate. And, they are both equally disposed to blame themselves for such a violation, but persist in the violation anyway due to a weakness of will. Yet, there is a difference between Alice and Beatrice that is relevant to whether their hypocrisy undermines their standing or not. Alice remains committed to trying to change her behavior and improve her strength of will, while Beatrice does not. Alice is *committed*, despite her akrasia, while Beatrice is *fatalistic*. Why would Alice's commitment to try to improve herself exonerate her hypocrisy if she does not actually change her behavior? I believe the answer lies in an acknowledgment of necessary human shortcomings and what it means to accept blame. When we are subject to blame, we are receiving a

request to acknowledge that our actions were wrong. Accepting blame entails a commitment not to engage in such behavior in the future. Yet, just in virtue of being human, we know that we will not always be able to guarantee our actions in the future. Sometimes, we simply slip up. A genuine acceptance of blame, then, doesn't require that we *never* do wrong again. It merely requires a commitment, a resolve to try not to do wrong to the extent that it is within our power.

Doing the right thing is hard. It often requires contravening our impulses and our self-interest. This means that in certain situations, we will get it wrong. But we do not think that this makes us bad people. What would make us bad is if we stop trying to get it right. Alice keeps trying to be a better person, and that is why we read her acknowledgment of her guilt as genuine. We are much less likely to think of Beatrice in the same way, because her guilt does not lead her to try to do better. The equality of persons means that everyone is equally accountable for their actions, barring extenuating circumstances. To be accountable is to make a commitment to improve, though making a commitment to improve does not necessarily entail actual improvement. If this is right, then there is a relevant difference between the committed akratic hypocrite and the fatalistic akratic hypocrite, which is that one of them genuinely responds to blame and the other does not. One is inappropriately partial to herself and the other is impartial in the relevant Kantian sense, and thus, one undermines her standing because she violates the governing principles of morality and the other does not.

To round out the analysis, consider how the exception-seeker undermines her standing. Unlike the other two wrongful hypocrites, who recognize that they are seeking a leg up and therefore undermine themselves by cheating, the exception-seeker undermines herself by ignorance. She really thinks that her violations of some norm are relevantly different than others in a way that makes them subject to account for their actions in a way she is not. She is wrong about this. She doesn't realize she is being inappropriately partial, but she is, and we can't allow her to operate as an equal player in blame until she corrects her ignorance. She undermines her standing by proving herself to be ignorant of the principles of the activity at which she pretends to participate. There may be no ill will, as in the case of the cheating hypocrites, but she still must be disqualified for not following the rules.

References

Bell, Macalester. 2012. "The Standing to Blame: A Critique". In *Blame*, edited by D. Justin Coates and Neal A. Tognazzini, 263–81. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199860821.003.0014.

Fritz, Kyle G., and Daniel Miller. 2018. "Hypocrisy and the Standing to Blame". *Pacific Philosophical Quarterly* 99, no. 1: 118–39. https://doi.org/10.1111/papq.12104.

Rawls, John. 1955. "Two Concepts of Rules". *The Philosophical Review* 64, no. 1: 3. https://doi.org/10.2307/2182230.

Roadevin, Cristina. 2018. "Hypocritical Blame, Fairness, and Standing". *Metaphilosophy* 49, no. 1–2: 137–52. https://doi.org/10.1111/meta.12281.

Rossi, Benjamin. 2018. "The Commitment Account of Hypocrisy". *Ethical Theory and Moral Practice* 21, no. 3: 553–67. https://doi.org/10.1007/s10677-018-9917-3.

Tosi, Justin, and Brandon Warmke. 2016. "Moral Grandstanding". *Philosophy & Public Affairs* 44, no. 3: 197–217. https://doi.org/10.1111/papa.12075.

Wallace, R. Jay. 2010. "Hypocrisy, Moral Address, and the Equal Standing of Persons". *Philosophy & Public Affairs* 38, no. 4: 307–41. https://doi.org/10.1111/j.1088-4963.2010.01195.x.

Gregory Antill (gantill@cmc.edu)

[Draft for NUSTEP 2019 – Please do not cite without permission]

Abstract: One reason the phenomenon of practical akrasia - cases where agents act contrary to their considered judgments about what they have most reason to do - has been of such lasting interest to philosophers is due to the thought that the possibility of practical akrasia can tell us something important about the structure of practical agency more generally. Just as the existence of practical akrasia has been treated as important evidence for the existence of our practical agency, the alleged *absence* of epistemic akrasia – cases where believers believe some proposition contrary to their considered judgments about what they have most reason to believe – has recently been marshalled as grounds for skepticism about the existence of similar forms of epistemic agency.

In this paper, I defend the existence of epistemic agency against such objections. Rather than argue against the claim that epistemic akrasia is impossible, I argue that the absence of epistemic akrasia is compatible with the existence of epistemic agency. The crucial mistake, I claim, is that skeptics about epistemic agency are failing to carefully distinguish between differences in the structure of believing and acting and differences in the structure of normative reasons to believe and normative reasons to act. I argue that differences of the latter sort can provide an alternative – and superior – explanation for the absence of epistemic *akrasia* than an explanation involving differences in our practical and epistemic agency.

I. Agency and Akrasia

One reason the phenomenon of practical *akrasia* – cases where agents act contrary to their considered judgments about what they have most reason to do – has been of such lasting interest to philosophers is because such cases are thought to hold the potential to tell us something important about the structure of practical agency more generally. The existence of practical akrasia has been treated as an important form of abductive evidence for the existence of some agential capacity, exercised when acting for reasons, whose failure or weakness makes it possible for agents to act contrary to their evaluative judgments about the force of their reasons. While agents may usually act as they judge they have most reason to act, the possibility of practical *akrasia* suggests that even in normal cases, there must be something further going on explaining why agents act in accordance with their reasons on those occasions, but not when they act akratically.[17]

[17] For a summary, see Stroud and Tappolet (2003):8-12. While there is some general consensus that *akrasia* reveals some important form of agency, there is widespread disagreement about what sort of capacity or mechanism the agency consists in. The significance of *akrasia* is perhaps put most explicit for those who hold that our agency consists in the exercise of some reflective capacity, as in Korsgaard (1996); Wallace (2001); Albritton (1985); or Normore (2007). But *akrasia* can also be used to help reveal the contours of our practical agency even for those in a broadly Aristotelian tradition who hold views associating our agency with the work of practical reason more directly such as Hieronymi (2009), Arpaly (2000), or Anscombe (1959). My discussion will not presuppose any view on what the relevant capacity is, or even, for that matter, that one accept the link between *akrasia* and agency at all. My central goal is rather methodological: I hope to show in this paper that even if one *did* hold that the

In contrast to the case of akratic action, there continues to be widespread skepticism about the existence or even the possibility of akratic belief – cases where believers believe some proposition contrary to their considered judgments about what they have most reason to believe.[18] Apparent examples of straightforward akratic believing seem much more difficult to come by than the seemingly abundant examples of akratic action. To many, the very possibility of such cases has seemed implausible; philosophers have sensed "a whiff of Moore's paradox" in the claim that, e.g., my evidence supports the conclusion that it is raining, but that I don't believe that it is raining.[19]

Just as the existence of practical *akrasia* has been treated as important evidence for the existence of our practical agency, this alleged *absence* of epistemic *akrasia* has recently been marshalled as grounds for skepticism about the existence of similar forms of epistemic agency.[20] If we exert our epistemic agency when we form beliefs on the basis of the reasons we judge to bear on believing, the argument goes, it should be possible for our agency to err, and so we should expect believers to occasionally fail to produce the requisite belief grounded in the recognized reasons. The alleged absence of any such agential breakdown appears to signal that there is no agency involved to break down in the first place.

Thus Adler (2002) argues for the conclusion that akratic believing is impossible, and claims that this conclusion "undermine[s]…attempts to assimilate theoretical [and] practical reasoning" and "reveals a fundamental divide between them."[21] Owens (2017) argues that "epistemic *akrasia* is impossible and that it's impossibility casts a shadow over the whole idea of doxastic control."[22] Setiya (2013) expresses "skepticism about epistemic agency," arguing that the impossibility of believing contrary to our judgments of what we most reason to believe helps "reveal a basic contrast" between what is involved in acting for reasons and believing for reasons. Whereas acting for a reason is the result of some rational causation which moves an agent from judgments about how she has reason to act to her acting for those reasons, "believing for a reason *reduces* to a conjunction of beliefs [the belief that p and the belief that r is evidence that p]. There is no further causality that connects them."[23]

In support for this deflationary account of believing-for-reasons, Setiya marshals precisely the sort of abductive argument for skepticism described above.[24] Says Setiya:

---

capacities which make *akrasia* possible are essential to agency, the absence of epistemic *akrasia* would *still* not provide reason to doubt the existence of our epistemic agency.

[18] See Adler (2002); Hurley (1993); Raz (2007); Pettit and Smith (1996); Owens (2017). For defense that Akratic belief *is* possible, see Scanlon (1998), Heil (1984); Borgoni and Luthra (2017). A few important clarifications about the possibility under debate: First, crucially, the sort of judgments being considered involve 'epistemic' reasons. There are few who deny that we might believe we have most pragmatic reason to believe some proposition, yet fail to believe it. Second, just as the akratic action must be intentional, the akratic belief must be attributable to the believer in the right way. There are few who deny that we might have recalcitrant beliefs – such as the phobic belief that spiders are dangerous – which are counter to our considered judgments. There is some debate about what this 'attributability' condition consists in. Owens (2017) holds that the akratic belief must be "formed freely and deliberately," Moran (2001) holds that the belief must be "avowed," Luthra and Borgoni (2017) hold that the akratic belief must be 'avoidable through the exercise of rational capacities.' In the proceeding, I will attempt to remain neutral on what sort of attributability akratic believing involves.

[19] Setiya (2013):197. I will discuss the details of such arguments against the possibility of akratic believing in greater detail in section III.

[20] See, e.g., Setiya (2013); Adler (2002); Owens (2017)

[21] Adler (2002):1-2.

[22] Owens (2017): 37.

[23] Setiya, (2013):196.

[24] In Setiya's paper, the alleged absence of akratic believing is just one among several important disanalogies between belief and action from which he argues for his deflationary conclusion. Other phenomenon include the

> it follows from your beliefs about the evidence that p that you believe it on corresponding grounds. What accounts for these phenomena if believing for a reason is not a mere conjunction of beliefs?[25]

As Setiya's paper reveals, this argument from *akrasia* is just one example of a larger methodological trend, starting with differences between how we believe and intend, and extrapolating from this to underlying differences between the structure of belief and intention. In this instance, the alleged absence of akratic belief is the grounds for an abductive argument against epistemic agency. If we never fail to believe what we judge ourselves to have most reason to believe, that is evidence that the transition from judgment to belief is unmediated. If we still insist that our beliefs are the result of our epistemic agency, we are now pressed to provide some further explanation of why this purported agential capacity never goes awry, producing cases of akratic belief.

Given the implausibility that humans exert a universally iron will over their epistemic agency so as to *never* be swayed or tempted to believe counter to their reasons, the lack of such agency may appear to be the best, and perhaps only, candidate explanation for the complete absence of *akrasia* in the doxastic realm. As Setiya puts the point in the passage quoted above: "what [else] could account for the phenomena?"

The rest of this paper is an attempt to provide a satisfactory answer to this question. I will argue that there are actually two possible kinds of explanations for restrictions on attitudes, one stemming from our agency with respect to that attitude, the other stemming from the environment in which the agent is situated. I will argue further that environmental restrictions can better account for the absence of *akrasia* in the doxastic realm than a lack of epistemic agency.

This study of akrasia will reveal an important cautionary lesson fo the comparative analysis of practical and theoretical agency more generally. When extrapolating from differences in patterns of acting and believing to differences in the structure of our agency, we must be careful to be sure that the differences really stem from differences in the relevant psychological structure of our practical and theoretical attitudes, not from differences the structure of the values and reasons in the normative domain which those attitudes are tracking.

II. Agency and the Environment

Suppose we encounter a car which, unlike other cars, never turns or swerves from its route. There are, broadly speaking, at least two potential sources of explanation for the absence of such activity. The first source is what we might call an *internal restriction*. The object might fail to exhibit certain activity because it lacks some of the internal structure required for the activity in question. If we see a car which never turns, for example, one explanation may be that it lacks the steering mechanism required for turning.

---

alleged absence of doxastic analogues to practical cases of rationalization, deviant causal chains, and choice without preference. While this paper will have implications for the force of his related arguments, the discussion of *akrasia* will not suffice to settle whether these other phenomena support Setiya's skepticism. For as I hope to show, one lesson of this paper is that each phenomenon may have different relevance, depending on the source of the disanalogy, and each and must be considered on a case by case basis.

[25] Setiya (2013):198.

The second source is what we might call an *environmental restriction*. An object might fail to exhibit certain activity because something about its external environment prevents the activity from occurring. Though the object has the necessary structural capacities to perform the activity, the exercise of the capacity is curtailed by the environment in which the object is situated. In our vehicle example, it may be that the course on which the car is driving has no side streets on which to turn. If a car were located on such a track, the complete absence of any turning activity would be no evidence for the absence of a steering capacity. Though the driver may retain the capacity to steer as she ordinarily would, this capacity is masked by the environment in which the car is situated: there are simply no alternative paths for the driver to steer the car on to.[26]

More generally, if we know that there exists an environmental restriction, the absence of activity no longer provides any strong abductive evidence for the absence of a steering mechanism. Were there such a mechanism, we should expect it to be masked by the external environment. Turning now to the case of belief: what are the possible sources of explanation for the absence of cases of epistemic *akrasia*?

One possible source would be the sort of internal restriction posited by Setiya. There could be some internal agential capacity exercised when a person acts for a reason, which is absent when a person believes for a reason. Because there is no capacity at work when an agent believes for reasons, the subject is unable to exercise that capacity to believe against what she judges she has most reason to believe.

What, if anything, would be the relevant environmental restrictions? The answer is less straightforward. There is no obvious analogue to the car's track for attitudes like belief. With some work, however, an analogy can be formulated. Beliefs and other attitudes held for reasons are often understood as *evaluative commitments*. Intentions are commitments about what to do, beliefs are commitments about what is so.[27] Just as cars are navigating the road, these evaluative commitments are attempting to navigate some normative terrain. In the case of commitments about what to do, we are attempting to evaluate actions as good or bad. In the case of commitments about what is so, we are attempting to evaluate propositions as true or false. Explanations which rely on differences between actions and propositions would be differences not in how we reason, so to speak, but in what we are reasoning about – differences in the external facts about true propositions and good actions which are the objects of our beliefs and intentions, rather than differences in the internal structure of believing and intending. Environmental restrictions, then, can be understood as restrictions about what we can believe or intend whose source lies in differences between the relevant normative structures of goodness and truth.

These normative restrictions are environmental in that they, like a physical environment, can mask the existence of internal agential capacities. This is perhaps easiest to see by illustration. To see such normative environmental restrictions at work, consider a case from the philosopher Joseph Raz concerning our capacity to act-for-reasons. According to Raz, there are a number of restrictions on when it is possible to perform certain actions for certain reasons. I can decide to go to one play over another for the reason that I love Sophocles, for example, but I cannot decide to have coffee over tea for the reason that I love Sophocles.[28]

Does this reveal some important difference in my capacity to act for reasons with respect to plays over beverages? Raz thinks no. In both my beverages and play choices, I am equally capable of being

[26] A related point is made by Lewis (1997). We cannot characterize capacities, or robust dispositions, in simple counter-factual terms, because the environment may ensure there is no counter-factual scenario where the disposition is actualized, though the individual possesses it.

[27] See e.g. Scanlon (1998); Moran (2001); Hieronymi (2005). While widely held, this view is not universally accepted. I will assume, rather than argue, for it here.

[28] Raz, (2001).

motivated by any reason I take to bear on the desirability of the activity in question. It is just that a love of Sophocles is relevant to the goodness of one play over another, but not to the goodness of one beverage over another, so there will be no such reason available to motivate me.

In this case, a certain absence (cases of deciding to get coffee for the reason that we love Sophocles) does not cast doubt on our agency with respect to beverage choices, because we have an explanation of the absence in terms of difference in normative environment. The absence is the result of differences in what is good about the kinds of actions being intended, not a difference in the agency exerted in intending to do the actions. It is, so to speak, a difference in the object of the attitudes – a difference between plays and coffee –rather than a difference in the attitudes themselves, which explains why I cannot perform certain actions out of a love of Sophocles.

Another example can be found in Schroeder (2007)'s "surprise party" cases. My love of surprise parties will never motivate me to head to a surprise party, even though my love of other activities will motivate me to go to the activity. This does not reveal any important limitation on the nature of my agency: my intentions are not somehow practically unresponsive to the fun of surprise parties. It is because of the nature of what's good about surprise parties: to be surprising, I can't be aware of them, so there will never be anything about the good situation for my regular internal agential structure to respond to. Such external features of the object of my intention provide environmental explanations for the absence of such attitudes.

While the above examples involve intentions with different objects, we can apply the same lesson to comparisons between intention and belief. Reasons for belief are truth governed. Reasons for action or intention are good-governed. When extrapolating from differences in when we act and believe to differences in the nature of action and belief, we should be careful to make sure the difference is in the acting and believing, not in the external facts about when reasons to believe and reasons to act are available.

This opens up at least the possibility of an environmental distinction which could explain the absence of epistemic *akrasia* without calling into question the existence of epistemic agency. In the following section, I will argue that if akratic belief is impossible, it is precisely this sort of environmental diagnosis which provides the most plausible explanation why. Insofar as we have a convincing argument that epistemic *akrasia* is impossible, the argument will have to rest on differences between the structure of goodness-given and truth-given reasons, not on differences between the structure of intention and belief. In particular, I will argue, recent objections to the possibility of epistemic *akrasia* rely heavily on the fact that two contrary propositions cannot both be true, in a way that two contrary actions might both be good.

III. The Arguments Against Akratic Belief

Those skeptical about the possibility of epistemic *akrasia* claim there is something paradoxical about the possibility of one's, e.g., simultaneously judging that the evidence supports the conclusion that it is raining, yet believing that it is sunny, in a way where it is not equally paradoxical to, e.g., judge that one's reasons count in favor of staying in to grade papers, yet decide to go get ice-cream in the park. In this section, I will survey the variety of objections offered by recent skeptics against the possibility of believing akratically by the same methods by which we achieve akratic action. Rather than focusing on evaluating whether such objections are successful, I will instead aim to show that each objection, successful or no, is best understood as grounded in environmental, rather than internal, limitations. Since these objections all rely on environmental limitations to reach the conclusion that epistemic *akrasia* is impossible, the impossibility of such *akrasia* will be an inappropriate starting point for arguments which attempt to move from the absence of epistemic *akrasia* to the absence of epistemic agency.

Following the general Davidsonian line on the origins of practical *akrasia*, recent skeptics begin with the observation that actions and their outcomes can be good in a variety of different ways – they might be moral, or pleasurable, or intellectually stimulating – and these different values can give rise to a variety of conflicting reasons for acting.[29] This variety of "goodness-related" reasons makes akratic action possible. When you act in a way that you judge you should not, you can still be responding to the appeal of some genuine good, even when you might have made an all things considered judgment that this good is outweighed by some other practical considerations. The more important reasons you have to grade your papers do not make the ice-cream taste any less delicious. And so the pleasure-related reasons you have can still make your akratic action intelligible, by showing how you see it as good or desirable in some respect, even though you think you have stronger reasons to stay in and grade.[30]

In contrast to the variety of sources of reasons for action, there is what Hurley (1993) has described as a "unity of reasons" for belief:

In the case of what should be done there may be conflict within an agent. There may be competing reasons conflicting for authority. But in the case of what should be believed, truth alone governs and it can't be divided against itself or harbor conflicts. It makes sense that something is, ultimately, good in some respects but not in others…in a way it does not even make sense to suppose that something is, ultimately, true in some respects but not in others.[31]

An action or outcome can be good in some respects, bad in others. But since truth, unlike goodness, is a univocal value, there are no multiple sources of value to provide us with countervailing reasons not to believe as we, all-things considered, judge we ought to do.

Following Hurley, recent skeptics of epistemic Akrasia such as Adler (2002), Raz (2007), and Owens (2017) have used this feature of truth to try to explain why it is impossible to believe akratically by the same method as we achieve akratic action.

The central thrust of their arguments are strikingly similar. Here is Owens (2017):

If one thinks the evidence establishes p, one must think that apparently countervailing evidence e' can be explained on the hypothesis that p and so provides no grounds for thinking not-p to be true. Should one nevertheless be swayed by the appearance of e', one is being swayed by a consideration whose probative force one can't acknowledge in judgment.[32]

From Raz (2007):

Epistemic reasons can conflict, but all of them are about the truth of the propositions for or against belief in which they are reasons. The weaker reasons are just less reliable guides to one and the same end. …Because there is no possibility that the lesser reason for belief serves a concern which is not served better by the better reason there is no possibility of preferring to follow what one takes to be the lesser

[29] Davidson (1969)

[30] Davidson's argument here assumes a background position that intentional action must be action done for some reason. In contrast, some action theorists, following Anscombe (1959), hold that some intentional actions, such a whims, can be intelligibly performed for no reason at all, so long as requests for reasons do not lack application.

[31] Hurley (1993): 133

[32] Owens (2017):45

reason rather than the better one. The possibility of [practical] *akrasia* depends on the fact that the belief that a practical reason is defeated by a better conflicting reason is consistent with belief that it serves a concern which the better reason does not, and which can motivate one to follow it.[33]

And Adler (2002):

When beliefs conflict, they weaken one another, since both cannot be true. When one belief is favored by the evidence, the disfavored belief evaporates, since it has been determined to be false. But when desires conflict, as with desires to pursue careers both in medicine and in ballet, the conflict need not, and typically does not, weaken either. When one desire is acted upon, the other retains a hold, experienced as regret…let us extend [this] observation from beliefs to the evidence or reasons for them. When evidence is adequate….then we accept or fully believe it. Consequently, and this is the crucial claim, previously conflicting evidence (i.e., evidence that supported a contrary of h) is nullified as undermining h.[34]

In each passage, the author trades off what he takes to be an important difference between reasons for belief and reasons for action: that while sufficient reasons for acting may outweigh the reasons for contrary actions, reasons for believing, when sufficient, undercut or nullify any possible reasons for believing contrary propositions. Because truth cannot contradict truth, if p is true then any true considerations which seemed to indicate that p was false must in fact be compatible with p after all. While ice-cream may still taste sweet, and so be good in some respect even though I know I should be grading, it cannot still be true in some respect that, e.g., it will rain, if the truth is that it will remain sunny outside. Thus my alleged evidence of rain (e.g. the dark clouds in the sky) cannot really be foreshadowing future rain, if there is no rain to foreshadow. They must actually be compatible with its being sunny out later. So even if I don't yet see how my apparently contrary evidence is compatible with what I believe to be true, I see that this piece of evidence must not, in fact, have the normative force which I had taken it to have.

Having lost my reasons for believing not-p, the argument goes, I have lost any possible motivating grounds for adopting the belief. My believing not-p is thus incompatible with my simultaneously judging that the evidence supports belief in some contrary proposition, and so paradoxical in precisely the way skeptics of epistemic *akrasia* had claimed.

With the argument for the impossibility of epistemic akrasia on the table, we can now assess whether, if it were sound, its conclusion would be an appropriate abductive starting point for skepticism about epistemic agency. I think, in this case, the answer is clearly 'no.' As in the case of *The surprise party* and *the Sophocles-Enthusiast*, it is clear, in the 'unity of reasons' argument, that it is the particular nature of truth and truth-given reasons, not the particular nature of belief, which explains why epistemic *akrasia* is impossible where practical *akrasia* is not. According to the skeptical arguments surveyed, beliefs and intentions are similarly structured in that the object of their respective evaluative commitments constrains the reasons for the attitude. They differ only in whether truth-given or goodness-given reasons for akratic actions or attitudes are available for the agent to hold. Akratic believing is impossible, according to the argument, not because we as believers lacks any capacity to be tempted by reasons for

[33] Raz (2007): 7
[34] Adler (2002): 6-7

believing propositions we judge we ought not believe, but because there will never be any such reasons to tempt us.

IV. Akrasia and the Environment

In the previous section, I have taken myself to show that insofar as we have good arguments which show that epistemic *akrasia* is impossible, these arguments rely on an external environmental source. Epistemic *akrasia* is impossible because of distinctive features of truth and truth-given reasons, not because distinctive features of the structure of believing.

Given the environmental source of the absence of epistemic *akrasia*, the absence cannot be grounds for doubting that we have epistemic agency. It would not reveal, as proponents have sometimes thought, a fundamental divide between the agential structure of our practical and theoretical attitudes. The absence of epistemic *akrasia* could be fully explained by as the result of those agential capacities being constrained, in the case of belief, by the believer's normative environment.

One might worry that such a constrained capacity to believe for reasons will constitute something of a pyrrhic victory. Perhaps our lack of epistemic agency is not a lack of some internal capacity, but rather just *consists in* our (in this case environmental) inability to do otherwise than what epistemic reason demands.

In this section, I will argue against this response by considering a variety of the Twin-Earth thought experiment. Imagine a world where the goodness of states of affairs happened to be structured in the same way that the truth of propositions is structured in the actual world. This would be a world with one univocal kind of good, or where the variety of goods aligned and so 'spoke with one voice' as truth does. Imagine, for instance, a world where pleasure is the only good, or where acting morally really was always in your own prudential self interest. To mimic the structure of truth in the actual world, our imagined world would also have to be a world where an action being good in some respect could not be bad in that same respect. Imagine, for instance, a world where every pleasurable activity in the short term was also in the interests of your long-term happiness. Finally, to mimic the structure of evidence in the actual world, our imagined world would have to be a world where having the most practical reason to act in one way means that any reasons an agent thought they might have had to act another way will be undercut rather than outweighed. We can imagine, for instance, that the most enjoyable activity among your options was always the only enjoyable activity. This would be a world where, if going to the park rather than the cinema would be more enjoyable, it will turn out that the cinema was closed anyway, so there was nothing enjoyable to regret missing out on when choosing to act as you had most reason to act.

Imagine we pluck a full practical agent from our world and place her in this new Panglossian world. What would happen to someone lucky enough to knowingly find herself in such an environment? There would never be any case where the reasons for two actions were counter-balanced, nor would there be any case in which an action one ought not to do had anything to be said for it, which an agent could take as her motivating reason to perform that action akratically. In such circumstances an agent might be no more able to act contrary to what she judges she has most reason to do than she would be able to believe contrary to what she judges she has most reason to believe. Though her will is just as susceptible to temptation as ever, her new environment is structured so as to lack any reasons to be tempted by, motivating her to act akratically.

But surely it would be strange to think that a previously free agent with a capacity to act at will, placed in such alternate circumstances but intrinsically unchanged, has somehow *lost* her agency or volitional capacities. We should not say she has suffered a change to the structure of her will. Rather,

what we should say is that her environment is structured in such a way which ensures that she will always exercise her will so that she acts in accordance with what she judges herself to have reason to do.

If this is right, we ought to say the same thing about ourselves as epistemic agents in the actual world. It is not that we lack epistemic agency, we just find ourselves in a normative environment which is structured in such a way that we will always exercise our epistemic agency to believe in accordance with what we hold ourselves to have most reason to believe.

IV. Conclusion

In this paper, I have defended the existence of epistemic agency against a threat to that agency, coming from the alleged absence of epistemic *akrasia*. Rather than showing that epistemic *akrasia* is possible, I have argued that the link between agency and *akrasia* has been overstated. This is because the possibility of *akrasia* rests not only on an individual's agency, but also on the environment in which she is situated. She must have some motivating reason for choosing to act contrary to her considered judgments. If her environment is such that she lacks any motivating reasons, *akrasia* may be beyond her, even though she has the agential capacities to act on such motivating reasons, were they available.

I have then surveyed the arguments for the impossibility of epistemic *akrasia*, and shown that they reach their conclusion by relying on just such environmental factors. Epistemic *akrasia* is impossible, opponents allege, not because believers are incapable of being motivated to believe akratically, but because they lack any available reasons to be motivated by. Given that epistemic *akrasia*, if impossible, is impossible because of environmental factors, the absence of epistemic *akrasia* is not actually grounds for skepticism about the extent or force of our epistemic agency.

References

Adams, Robert (1985). "Involuntary Sins," *Philosophical Review* 94 (1):3-31.

Adler, Jonathan (2002). "Akratic Believing?" *Philosophical Studies* 110 (1):1 - 27.

Albritton, Rogers (1985). "Freedom of the Will and Freedom of Action." *Proceedings and Addresses of the American Philosophical Association* 59 (2):239-51.

Anscombe, G. E. M. (1959). *Intention*. Harvard University Press.

Arpaly, Nomy (2000). "On Acting Rationally Against One's Best Judgment." *Ethics* 110 (3):488-513.

Davidson, Donald (1969). "How Is Weakness of the Will Possible?" In Joel Feinberg (ed.), *Moral Concepts*. Oxford University Press.

Flowerree, Amy (2017). "Agency of Belief and Intention" (2017) Synthese 194 (8):2763-2784

Heil, John (1984). "Doxastic Incontinence." *Mind* 93 (369):56-70.

Hieronymi, Pamela (2005). "The Wrong Kind of Reason" *Journal of Philosophy* 102 (9):437-457

Hieronymi, Pamela (2009). "The Will as Reason." *Philosophical Perspectives* 23 (1):201-220.

Hurley, Susan (1993). *Natural Reasons*, Oxford: Oxford University Press.

Korsgaard, Christine (1996). *The Sources of Normativity*. Cambridge University Press.

Lewis, David (1997). "Finkish Dispositions." *Philosophical Quarterly* 47 (187):143-158.

Luthra, Yannig and Borgoni, Christina (2017). "Epistemic *Akrasia* and the Fallibility of Critical Reasoning." *Philosophical Studies* 174 (4):877-886.

Moran, Richard (2001). *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton University Press.

Normore, Calvin (2007). "Freedom, Contingency, and Rational Power." *Proceedings and Addresses of the American Philosophical Association* 81 (2):49 - 64.

Owens, David (2017). *Normativity and Control*, Oxford: Oxford University Press: 37-51

Pettit, Philip and Smith, Michael (1996). "Freedom in Belief and Desire." *Journal of Philosophy (93)9 429-449.*

Plato (1961). *Protagoras*, in *The Collected Dialogues of Plato*, E. Hamilton and H. Cairns (eds.), Princeton: Princeton University Press, pp. 308-352.

Raz, Joseph (2007). "Reasons: Practical and Adaptive." SSRN elibrary.

Raz, Joseph (2001). *Engaging Reason*. Oxford University Press.

Scanlon, T.M. (1998). *What We Owe to Each Other.* Cambridge: Harvard University Press.

Schroeder, Mark Andrew (2007). *Slaves of the Passions*. Oxford University Press.

Setiya, Kieran (2013). "Epistemic Agency: Some Doubts" *Philosophical Issues* 23 (1):179-198.

Stroud, Sarah and Tappolet, Christine (2003). "Introduction" in Stroud, Sarah & Tappolet Christine, (eds.) (2003). *Weakness of Will and Practical Irrationality*. Clarendon Press.

Wallace, R. Jay (2001). "Normativity, Commitment, and Instrumental Reason." *Philosophers' Imprint* 1:1-26.

The Importance of Logically Complex Actions

Andrew Flynn

1. Introduction

Imagine the following, bizarre hypothetical: you wake up one morning and find to your horror that your friends are all engaged in the weirdest activities, activities you've never even heard of before. Paul is having-beer-or-wine-with-dinner like there's no tomorrow. Jack has started off on an epic session of buying-rice-milk-if-there's-no more-almond-milk-left-at-the-grocery-store. And, Heather is in the midst of finishing-that-novel-she-started-last-weekend-unless-something-more-interesting-catches-her-eye-in-the-bookstore. What to make of this strange state of affairs? Unfortunately, I'm not around to help you out yet—I'm off in my library carrel, defending-or-defeating-the-philosophy-of-G.E.M.-Anscombe. *Strange goings on!* – to quote Donald Davidson.

Quite so. In this paper, however, I want to defend the existence of such strange goings on. Or, more precisely, I want to argue that recent discussions of the imperfective aspect and the structure of action by philosophers drawing on Anscombe's work imply that such strange goings on exist. This is not, as it turns out, a trivial result. Philosophers of action divide, roughly, into two camps: those who work in conceptual terrain shaped primarily by Davidson and Michael Bratman (henceforth traditionalists) and those who work in conceptual terrain shaped primarily by Anscombe (henceforth Anscombians).[35] These philosophers disagree about a whole host of issues—whether action explanation is a species of efficient causal explanation, whether intentions are mental states at all, whether self-knowledge is importantly connected with action—and the dialectic between the two camps is difficult to navigate, not insignificantly due to the fact that it is underdeveloped in the current literature. However, one subject that the Anscombians have focused on, but which has really not appeared on the radar of traditionalists, is the means-end structure of actions in their progressive unfolding over time.[36] In this paper, I hope to show that numerous arguments in the traditionalist literature, due to inattention to this feature of actions, tacitly

[35] The works I am thinking of here are the essays collected in the first section of Donald Davidson's *Essays on Actions and Events* and Michael Bratman's *Intention, Plans, and Practical Reason*—as well as his subsequent amendments made to the basic theory presented in that book—which have had a pervasive effect on how action theory is practiced in contemporary philosophy. On the Ansombian side, the *locus classicus* is Anscombe's *Intention*, but two recent attempts to revive Anscombe's work seem especially important: *Reasonably Vicious* by Candace Vogler and *Life and Action* by Michael Thompson. This is not to say that these are the only significant fault lines in action theory. For instance, see Setiya 2010 for a discussion of the numerous views that one could take on the nature of intentions, which contrasts the predominant traditionalist view that intentions are mental states with the Anscombian claim that they are not, but also considers philosophers, like George Wilson, who dissent from the predominant view, but not on explicitly Anscombian grounds. (See Wilson 1989.) Also, there are heated debates within each roughly defined camp, although they will tend to be debates that accept as given certain controversial theses, and then proceed to debate the implications or nature of those theses. So, for example, traditionalists take it as given that intention is some sort of mental state, but have fierce debates about the nature of the rational norms that govern that mental state. See, for example, Ferrero 2012a and Bratman 2012a. Anscombians, on the other hand, take it as given that a certain type of robust self-knowledge that Anscombe called "practical knowledge" is importantly connected to intentional action, but debate the nature of that knowledge. See, for example, Thompson 2011, Haddock 2011, and McDowell 2011. Finally, these are only rough constellations of positions that happen to hang together in the current literature. Some philosophers have mixed and matched positions in interesting ways. Kieran Setiya and J. David Velleman, for instance, have attempted to accommodate Anscombe's points about the centrality of self-knowledge to action while maintaining that intention is a mental state. See Setyia 2007, Setiya 2008, Setiya 2009, Setiya 2011, and Velleman 1989.

[36] Michael Bratman might take issue with being labeled a paradigmatic traditionalist, because he denies that he is guilty of an "eye-blink-like," atomic view of action that Michael Thompson attributes to traditionalists. (See Bratman 2012b, pg. 8.) I think his denial is right; Bratman's work is attuned to the temporal dimensions of agency. But, as I will argue in §3.3, Bratman is not sufficiently attentive to the means-end structure of actions that Anscombians take to be revealed in the unfolding of those actions over time.

assume that actions with what I will call a *logically complex* structure—that is, actions that involve means taken to logically complex ends—do not exist. This assumption leads these philosophers to adopt less plausible positive positions than they are in fact entitled to. I hope, then, that this paper shows that traditionalists ignore structure and aspect at their peril.

The paper will proceed as follows. In §2, I will introduce theses about aspect and the structure of actions commonly espoused by Anscombians. Then, I will connect these features to the role that the practical syllogism plays in Anscombe's theory. I will argue that these theses provide conceptual space for actions that have logically complex structures, and demonstrate the importance of this fact by considering a worry raised by Kieran Setiya. In §3, I will consider three case studies to show that the traditionalist literature has operated in circumscribed conceptual space as a result of a failure to appreciate the possibility that actions might have logically complex structures. Specifically, I will demonstrate that a number of arguments made by Bratman, Hugh McCann, and Richard Holton urging either the acceptance of mental states weaker than intentions or the loosening of the rational norms governing intentions also tacitly assume that actions cannot have logically complex structures. In §4, I conclude with some brief observations about the importance of exchanges between traditionalists and Anscombians.

2. The Importance of Anscombe

2.1. Introduction

In this section, I want to do the following. First, I will briefly sketch two features of recent work on Anscombe which have been stressed: aspect and structure. Then, I will connect these two features to a topic which has received less attention: Anscombe's account of the practical syllogism. I will use these point to argue that actions themselves may be conceived as having logically complex structures. Finally, I will engage with a worry that Kieran Setiya raises as a way to motivate my further criticisms of Michael Bratman, Hugh McCann, and Richard Holton.

2.2. Aspect and Structure

In recent attempts to highlight the importance of Anscombe's work, Anscombians have pointed to two features of intentional action that have gone missing from current discussions: aspect and structure.[37] Traditionalists, Anscombians complain, treat intentional actions as though they were essentially point-like, non-enduring primitives: the flipping of light-switches and the pulling of triggers.[38] But, nothing could be further from the truth. First, actions normally unfold over time, and so admit of what linguists refer to as aspectual distinctions. That is, they can either be completed or in progress towards completion. Actions described in the past tense admit of both imperfective and perfective aspect: "I was walking to the beach" and "I walked to the beach," respectively. Actions described in the present tense only admit of

[37] Some philosophers make much of the distinction between intentional actions and "plain-old" actions. See Vogler 2009 and Ford 2011, referencing Velleman and Frankfurt, among others. This distinction, however, does not play a role in the argument of this paper.

[38] Referring to his own Anscombian theory of intentional action, Michael Thompson writes: "That such a position seems strange... is in part a consequence of received conceptions of intentional action itself, above all, of the tendency of students of practical philosophy to view individual human actions as discrete or atomic or pointlike or eye-blink-like units that might as well be instantaneous for all that it matters to the theory" (2008, 90-1). Candace Vogler concurs: "For all that...'intentional action' functions as a kind of unanalyzed primitive in contemporary work" (2001, 45). On the relevance of these complaints to Bratman's work, see footnote two.

imperfective aspect—"I am walking to the beach" – since if an action is currently taking place, it has not yet finished unfolding.[39]

Second, it looks like actions, considered in their unfolding in the imperfective aspect, have rich internal structures. Consider an instance of cake baking: this is not some unanalyzable, instantaneous occurrence. Rather, intentional actions are, the Anscombians point out[40], means-end structured events, where the end is some state of affairs that an agent aims to produce and the means are other intentional actions that an agent performs in virtue of the fact that she takes them to be productive of that state of affairs.[41] So, an instance of cake baking involves, roughly, an agent's aiming to produce a cake, and then performing other intentional actions—breaking eggs, mixing batter, pre-heating the oven—in virtue of the fact that she takes them to be productive of a state of affairs in which a cake exists.[42] This account of structure makes sense of why instances of identical action types might nevertheless also be instances of other, different action types unfolding. The smaller intentional actions are not made fully intelligible until they are placed within the means-end structure of the larger intentional action of which they are parts.[43,44]

---

[39] The most important recent discussion of aspect is chapter eight of Thompson 2008. Thompson is drawing on chapter eight of Kenny 1963, Mourelatos 1978, Mourelatos 1993, Galton 1984, and chapter four of Vendler 1967. See also Frey 2012, pgs. 11-15 and Moran and Stone 2011, pgs. 47-55. Anscombe does not spend much time discussing aspect in *Intention*, although she makes remarks about the progressive nature of actions in §23, p. 39. And, as Thompson notes, Anscombe's writing offers a sharp contrast to Davidson's in the following respect: Anscombe's examples appear in the imperfective, whereas Davidson's appear in the perfective. (See Thompson 2011, 203-4.)

[40] For the most developed discussion of the means-end structure of actions, see chapter six of Vogler 2002, especially pgs. 127-35. See also Thompson 2008, pgs. 86, 93-4, and 106-12, and §12, 22, 23, and 26 of Anscombe 2000. In addition, see Frey 2012, pgs. 15-27. Sergio Tenenbaum's "Policy as Action Model" is also quite similar to Anscombian accounts of the structure of actions; see Tenenbaum 2010 and Tenenbaum 2012. Also, Ferrero 2012b offers an interesting, traditionalist discussion of closely related issues. I say "point out" rather than argue, because Anscombians have been responding to a lack of discussion of structure at all in the literature, rather than disputing some other account of structure on offer.

[41] This account of the structure of intentional actions raises some obvious worries about regress: if intentional actions are made up of intentional actions, then it looks like we've got intentional actions all the way down. Thompson, it seems, accepts as unproblematic an infinite downwards regress of intentional actions. (See Thompson 2008, pgs. 106-12.) Vogler admits that we may bottom out with some intentional actions that are in fact unanalyzable primitives, but points out that these are extremely marginal instances of action. (See Vogler 2002, pg. 257n18.) See also Millgram 2012 for discussion of the differences between Thompson and Vogler.

[42] Means and ends seem to be invoked as primitive notions in the recent Anscombian literature, but Anscombe seems to have understood the relation as I lay it out here: "That is to say: the future state of affairs mentioned must be such that we can understand the agent's thinking it will or may be brought about by the action about which he is being questioned" (2000, 35). The notion of an agent performing some action in virtue of taking it to be productive of some state of affairs invoked here is rough and intuitive, but I think that this is all that is needed for the purposes of my paper; again, to invoke Anscombe: "I shall not try to elaborate my vague and general formula, that we must have an idea how a state of affairs Q is a stage in proceedings in which action P is an earlier stage, if we are to be able to say that we do P so that Q. For of course it is not necessary to exercise these general notions in order to say 'I do P so that Q'. All that is necessary to understand is that to say, in one form or another: 'But Q won't happen, even if you do P', or 'but it will happen whether you do P or not' is, in some way, to contradict the intention" (2000, 36).

[43] I am framing the issue here slightly differently than it sometimes appears in the literature. I am talking *metaphysically*, as it were, about intentional actions that are parts of a larger intentional action in virtue of being means to some end that is definitive of that action. Anscombians—and Anscombe herself—frequently talk more *linguistically* about a series of nested descriptions under which an agent is acting. (See, for example, Thompson 2008 and Anscombe 1979.) I've decided to do this, because it seems to me that fully understanding what Anscombians mean when they talk about acting under a particular description or series of descriptions is tightly bound up with controversial and difficult theses about practical knowledge—theses about which many of the philosophers I will discuss in the second half of this essay are skeptical. For my purposes, I don't see that anything is lost by avoiding this question.

[44] The Anscombian account of structure should not be confused with an account of action individuation. Confusion might occur because Anscombe discusses action individuation in the context of means-end structure in §26 of *Intention*.

2.3. Anscombe on the Practical Syllogism

Once we have the picture of intentional action sketched in the previous section, the following point becomes important: an agent might set ends that are logically complex. She might, that is, plan around reading a book this evening, *if* there's nothing on television. But, this seems banal. Standardly, philosophers of action will understand agent in this situation as forming an intention for the future with complex content. Nothing seems to follow from these basic points about time and means-ends reasoning about the logically complex structure of actions.

We will return to a worry like this at the end of this section. However, to see what it would amount to for there to be logically complex actions, I want to turn I want to pursue the question this way: according to Anscombe, following Aristotle, an action is the conclusion of the practical syllogism — of a bit of practical reasoning.[45] So, let's take the following: "I'll read a book tonight, if there is nothing on the television." I want to think about how this could, on Anscombe's view of the practical syllogism, count as a conclusion of such a syllogism. It will turn out, I think, that, once we think this through, we will see that one should not contrast Anscombe's view with Davidson-inspired views in the way that Kieran Setiya does. However, we will also be able, I'll contend, to draw some important lessons for action theory in general.

What is the practical syllogism, or practical reasoning, according to Anscombe? It is a form of reasoning that starts from something wanted — from an end that an agent has — and concludes in an action. The reasoning moves from the end to specify what means are needed to bring the end about in such a way that an action currently available to the agent may take place. The reasoning goes well in a way that can be analogized with theoretical reasoning but which is not the same as theoretical reasoning. In theoretical reasoning, one moves from premises to a conclusion and the reasoning has good form if a truth-preserving pattern obtains between the premises and the conclusion. In the case of practical reasoning, the reasoning has good form if the goodness of the end is preserved through the means specified to the action which is the conclusion. (This statement is quite simplistic — we'd have to say more about the more complicated versions of both forms of reasoning — but it will do for the purposes of my paper.)

What sort of thing wanted, or what sort of end, would be specified by the conclusion: "So, I'll read a book tonight, if there's nothing on the television"? Let's say, I want to have a bit of fun tonight. That's my end: having a bit of fun. I might, however, just get a straight out action as a conclusion, specifying how I'm going to bring that end about: I might take it that watching a bit of television would be fun, so I'll watch some television. So, to get a natural story of how I might conclude with a logically complex action description, we'll need to bring in some more considerations which make the case more complex. Suppose, then, that I see that there are a number of ways to have some fun tonight that occur to me. Specifically, there are two ways that occur to me. I might finish up Ada Palmer's novel *Too Like the Lightening*. (I'm nearing the end and it is getting exciting.) However, really I know that I'll be tired at the

---

In one of the only non-Anscombian discussions of these mereological issues related to action of which I am aware, Sara Rachel Chant's "Two Composition Questions in Action," the author, so far as I can tell, straightforwardly confuses the two issues. Chant chooses to focus on mereological issues related to collective instead of individual action, because she takes the question of composition in individual action to be identical with the question of action individuation, and she takes action individuation to have been worked to death in literature. (See Chant 2010, pgs. 28-30.) But, it seems to me that these questions are pretty clearly distinct: offering an account of why pulling a trigger is identical to assassinating President Lincoln is different than explaining why buying a gun and deciding the best time to strike—two clearly non-identical intentional actions—are both parts of plotting to assassinate President Lincoln.

[45] In addition to *Intention*, see Wiseman 2017 and Vogler 2001.

end of a long day of explaining the *Nichomachean Ethics* to perplexed undergraduates. I want to have fun because I need a break to recharge myself. Reading Ada Palmer is not exactly like work — it's not like trying to understand the argumentative structure in the seventh chapter of Stephen Engstrom's *The Form of Practical Knowledge.* But I read all day many days for work, and however relaxing it would be to read a good sci-fi novel, I kind of just want to watch something fun. However, I also know that I find it frustrating to surf around Netflix endlessly and end up watching some episodes of the X-Files that I've already seen. I'd rather just read the book I know I'm enjoying, unless something really good catches my eye quickly. So, I conclude a version of what we started trying to theorize: So, if there's nothing that catches my eye on Netflix, I'll finish up *Too Like the Lightening* this evening.

2.4. Logically Complex Actions

Once we have the picture of the practical syllogism just sketched, we can notice the following: ends that are best described in logically complex ways, or intentions which are best understood as having logically complex content, are actionable in ways that are no different from simple, present directed intentions.

for instance, an agent might set the end of buying rice milk at the grocery store *if* there is no almond milk left. Or, an agent might set the end of having *either* beer *or* wine with dinner. These ends involve complex amalgamations of states of affairs. In the first case, it seems that the agent is aiming at producing one state of affairs given certain conditions obtaining, or some other state of affairs otherwise. In the second case, the agent is aiming at producing either of two states of affairs. And, the agent can take means to such ends. Means to having beer or wine with dinner, for instance, will be intentional actions that are productive of *either* state of affairs obtaining—that is, *either* a state of affairs in which she's consumed a glass of wine with dinner *or* a state of affairs in which she's consumed a glass of beer with dinner—and performed in virtue of that fact. These will be intentional actions that are neutrally productive of these two states of affairs—actions, like driving to the liquor store, which would equally further the agent's progress towards both those states of affairs—or are productive of just having beer with dinner while not making it prohibitively costly to have wine with dinner, and vice versa. (For example, the action of pouring wine into a carafe is not directly productive of a state of affairs in which an agent has beer with dinner, but, given an agent's normal background considerations, pouring wine into a carafe would not make it prohibitively costly to have beer with dinner—the way, say, spending all of her money on wine would—and an agent who had set the end of having beer or wine with dinner might perform it in virtue of this fact.)

When an agent is taking the means to such an end, what she is doing has the structure of a larger action unfolding, as per the Anscombian account. I will refer to such means-end structured events as having a logically complex structure, because they involve taking means to logically complex ends. For the rest of the paper, I will also focus primarily on a specific type of logically complex structure—disjunctive structure—since this will be most important for my criticisms of the traditionalist literature, although it seems to me that what I say should generalize to other instances of logically complex structure.[46]

[46] It seems that any expression of intention with a logically complex object that can be cashed out in terms of means taken in virtue of the fact that they would allow for some amalgamation of states of affairs to come about can be captured by my account. *Prima facie*, it seems like conditionals, negation, and conjunction should be similar to disjunction. Regardless, disjunction is so widespread, as this paper will show, that even if my account only captured actions with disjunctive structures, it would still be philosophically interesting.

2.5. Importance of Logically Complex Structure

To this point, I have argued that on the Anscombian account of structure, actions may have logically complex, specifically disjunctive, structures, and that many ordinary actions in fact exhibit such structures. This is an interesting result in itself, since disjunctive structure has not been discussed in the literature.[47] However, in the rest of the paper, I want to make a much stronger claim about disjunctive structure: there is good evidence that traditionalists have been working in circumscribed conceptual space as a result of their failure to appreciate the structure of actions that appears in the imperfective aspect. I will show that numerous arguments made by traditionalists are, in light of Anscombian points about structure, seriously flawed, and it seems likely these flaws stem from a failure to pay attention to the structure revealed in the imperfective aspect.

In what follows, I will consider a number of argumentative moves made by traditionalists, and argue that in light of the Anscombian account of structure and the conceptual space it provides for actions with logically complex structures, these moves are implausible and unmotivated. Further, I will argue that in each case the argumentative moves in question *are* plausible and motivated, however, given the implicit assumption that actions, in their progress towards completion, only involve means taken simply to produce the state of affairs that is eventually realized in the perfective—the state of affairs the agent counts as having intentionally produced when the action is completed—and not means taken to a disjunctive or otherwise logically complex list of states of affairs. (That is, for instance, in the case of an action which in the perfective is an instance of "buying a bottle of Pinot Noir," where the agent has intentionally produced a state of affairs in which a bottle of Pinot Noir is bought, that action in its unfolding only ever involved means taken simply to produce a state of affairs in which she'd bought a bottle of Pinot Noir, and not some logically complex end, like buying a bottle of Pinot Noir, *or* Pinot Grigio, *or*....) Indeed, in each case I think that it is very hard to motivate the philosopher's argumentative move *unless* this is assumed, strongly suggesting the traditionalists are tacitly relying on this assumption.[48] This assumption is false given that actions can have logically complex structures, but it would be an easy assumption to make if one primarily thought about actions in the perfective aspect. Since structure does not appear in the perfective aspect and since we almost never talk about completed actions in logically complex terms, one might easily assume that at every point in the action's progression towards completion, the agent performing it was always aiming to produce just the state of affairs that is ultimately produced. Not only do the traditionalist's arguments fail, then, but they probably fail due to reliance on an assumption that is held only because of an inattention to aspect and structure. Even if the reader is not convinced by these diagnostic inferences, though, I hope that it will be clear that in light of the Anscombian account of structure, numerous traditionalist arguments are insufficient. Anscombian complaints do have some bite, then, on the traditionalists' own terms, and so traditionalists cannot shrug

[47] See Ferrero 2010, pgs. 1-2, which notes that it is generally assumed that actions are not themselves "disjunctive."

[48] There is conceptual space for a related, weaker assumption: although the state of affairs that an agent aims to produce may grow more specific over time, an agent must always be taking means simply to produce a single state of affairs, not to produce any of a number of states of affairs. In each case discussed, these assumptions have the same result. Each case involves an agent taking means to two distinct states of affairs that don't, in the example provided by the philosopher, grow in specificity over time. The philosophers might, as I suggest, be assuming that actions can only be made up of means taken simply to produce the state of affairs that ends up obtaining when the action is completed. Or, they might be assuming that actions can only be made up of means taken simply to produce a single state of affairs, and that although that state of affairs might in principle become more specific over time, the only candidate for that single state of affairs in the cases in question is the state of affairs that ends up being realized when the action is completed. I discuss the first assumption for the sake of simplicity, but if you think it is more charitable to hold the second assumption, any of the arguments could easily be run using it. This appears to exhaust the conceptual space for denying that actions have logically complex structure.

off structure. To warm up, consider a worry raised by Kieran Setiya, about how to understand the way in which the guises of intention should be unified. This problem, a statement of which opens Anscombe's *Intention*, starts from the observation that we talk about intention in three ways. First, we express prior intentions for future actions. ("I intend to go to the store tomorrow.") Second, we discuss intentional actions. ("I'm intentionally taking a walk.") Third, we offer intentions with which our actions are done. ("I'm taking a walk with the intention of going to the store.") Anscombe thinks that *intention* is plainly not equivocal, and so we've got a puzzle: what unites these three disparate uses?[49] There has been broad agreement amongst action theorists about the importance of this puzzle, although the dominant way of attempting an answer, the one favored by traditionalists, is not the one that Anscombe seems to have favored. Traditionalists have taken prior intention to be a mental state and attempted to make sense of how this mental state is connected to intentional action and intention-with-which. Anscombe scarcely discusses prior intention in *Intention*, but Anscombians have essentially wanted to reverse the traditionalist procedure and analyze prior intention in terms of being, in some sense, already embarked upon the early stages of intentional action itself.[50]

This is a novel strategy which challenges a mode of proceeding that has essentially been orthodoxy in action theory since Davidson's later work.[51] But, in contrast to the traditionalist view, its details have not been worked out and it has lots of potential problems. Kieran Setiya states one of them as follows:

> It is a problem for the theory of intending as being embarked on intentional action that these objects can be logically complex. I intend not to be hit by a car as I walk home. I intend to drink wine or beer with dinner. I intend to read a book tonight if there's nothing on the radio. In none of these cases can we say, without contrivance or difficulty, what action I am now on the way to performing. Until it is supplied with an account of these cases…the theory of intending as being embarked on intentional action remains incomplete. (Setiya 2010)

When an agent forms a prior intention with—to continue this essay's example—a disjunctive object, the normal purpose of forming such an intention, on any account of what prior intentions are, is to keep the agent's options open for the time being. Normally an agent will form the prior intention to have wine or beer with dinner when she wants to keep open the option of having either wine or beer with dinner and plan the rest of her activities accordingly.[52] This will require her to take means to the end of *either* having wine or having beer with dinner, performing intentional actions in virtue of the fact that they are neutrally productive of either a state of affairs in which she has beer or a state of affairs in which she has wine, or productive of just one of those states of affairs but without making the other prohibitively

[49] See Anscombe 2000, pgs. 1-2.

[50] See Setiya 2010 for a detailed taxonomy of solutions to this problem. I think that some Anscombians would be unhappy with Setiya's way of framing the issue. As Moran and Stone explain, it is not just that Anscombe opposed the particular way of connecting the three notions that became orthodox in action theory. Rather, Anscombe was opposed to what they call "connective strategies" in general, strategies which take there to be three guises that need connecting. Anscombe thought that all of our expressions of intention were ways of picking out different instances of a single underlying form. "Anscombe's aim is to exhibit the unity of intention directly," they write, "by subsuming the three divisions under a single form" (44). (See also Vogler 2009.) However, I don't think that much hangs on this; the response that I give to Setiya's worry is to argue that prior intentions can have disjunctive structure, just as actions can. Whether we conceive of the Anscombian position as holding that prior intention involves already being embarked upon action, as Setiya explains it, or as holding that that prior intention is structurally of a piece with action, which seems closer to the way that Moran and Stone put things, does not affect whether its structure can be disjunctive.

[51] See "Intending" in Davidson 1980.

[52] See Ferrero 2010 for an account of disjunctive intentions from a traditionalist perspective.

costly. But, as I've already discussed at length in this essay, this is just to be, on the Anscombian account of structure, performing an action that has a disjunctive structure.

Of course, if the Anscombians are right and prior intentions involve being already embarked upon intentional actions, then the agent is performing an action that will eventually be, in the perfective, an instance of having beer with dinner, if that's what happens. But, that doesn't mean that the action, throughout its unfolding, involves only means taken to the end of having beer with dinner, as the earlier discussion of structure showed. And, although it is true that at the point in time at which the agent is carrying out an action with disjunctive structure, it is not determinate whether the action that is unfolding will—in the perfective—be an instance of having wine or an instance of having beer, this is also true when an agent is taking the constitutive means of many actions. If it is supposed to be problematic that an agent can be embarked upon an action when it is not yet determinate what that action will be an instance of in the perfective, then this is a worry that Setiya himself needs to grapple with just as much as the Anscombians.

While the Anscombian may encounter some problems in denying that prior intentions are mental states, Setiya's problem is not one of them. So why does Setiya think that it is hard to make sense of what the agent is doing without "difficulty" or "contrivance"? It is very hard to motivate his worry, unless one assumes that actions can only ever be made up of means taken simply to produce the state of affairs that the agent has intentionally produced when the action is completed. But, with this assumption in place, Setiya's worry suddenly becomes very pressing. In this case, it can't be that the agent is already embarked upon having beer with dinner, if that's what happens, because she is not taking means simply to produce a state of affairs in which she's consumed a glass of beer. Yet, that is what ends up happening, so for the Anscombian account of prior intention to work, it looks like that's the end the agent needs to be taking the means to. Perhaps, one might want to claim that an agent was already embarked upon *both* having beer and having wine with dinner, but only completed one. But, this seems wildly implausible: it fails to capture the exclusivity of the disjunction and seems to entail that there is a failed or abandoned attempt at having wine with dinner where there appears to be none. So what is the agent supposed to be embarked upon?

Given the Anscombian account of structure, however, the assumption is false. Setiya's objection either ignores the possibility that actions may be logically complex in their unfolding, or begs the question against the Anscombian by assuming that actions cannot be logically complex in their unfolding.

3. Three Case Studies

3.1. Michael Bratman's Endeavorings

Next, let's consider an argument by Michael Bratman.[53] Bratman offers the following thought experiment: an ambidextrous video game player is playing two of the same video game machines, one with each hand. The game involves shooting at a target. But, the video game machines are hooked up such that, if she hits either target—target A or target B—both machines will shut down. The most effective means of hitting one of the targets, though, is shooting at target A and shooting at target B, so this is what the video game player does, knowing that she can only hit one. And, when she does hit target A, she counts as hitting target A intentionally.[54]

[53] See Bratman 1987, pgs. 111-27.

[54] This final assumption, that when the agent hits target A, she counts as having done so intentionally, is widely accepted. See Setiya 2010. However, I think the Anscombians would reject it, since they tie intentional action so closely to self-

Bratman takes this to be a counterexample to a common assumption about the relationship between intentions and intentional actions, the "Simple View." The Simple View holds that whenever an agent ϕs intentionally, she has the intention to ϕ. In Bratman's video game example, the agent hits target A intentionally, so if the Simple View is correct, then the agent has the intention to hit target A, and this explains her shooting at target A. But, the agent is behaving identically with respect to the end of hitting target B—that is, she is shooting at target B with as much effort and skill as she is shooting at target A—so if her shooting at target A is explained by her intention to hit target A, it seems like her shooting at target B should be explained by an intention to hit target B. If the Simple View is correct, then, in the video game example the agent has the intention to hit target A and the intention to hit target B.[55]

Since the agent knows that it is impossible to hit both targets, however, this conclusion violates the principle of strong consistency, that an agent's intentions should be synchronically consistent with all of her beliefs. (I.e. it should be possible for the agent's plans to be successfully carried out if all of her beliefs are true.) And since the principle of strong consistency is required for intentions to serve as the planning states Bratman thinks that they are, he takes this to be a very bad conclusion. In the face of this argument, Bratman urges us to give up the Simple View.

In response, Bratman holds that there is a weaker mental state, endeavoring, that makes sense of the rationality of the video game example. Endeavoring is a guiding desire that is not governed by the norm of strong consistency. So, in cases like the video game example, the agent hits target A intentionally, but, contrary to the Simple View, she never intended to hit target A. She endeavored to hit target A and endeavored to hit target B.[56]

Bratman's rationale for positing this mental state is little more than the need to make sense of the rationality of examples like the video game example.[57] However, in light of the Anscombian account of structure, the video game example seems to be an example of an action with disjunctive structure. The video game example seems structurally identical to having wine or beer with dinner, where the agent did things that were productive of having wine—pouring wine in a carafe—and things that were productive of having beer—putting beer in the fridge—but in virtue of the fact that they were productive of either state of affairs. Here, the agent has the end of *either* hitting target A or hitting target B, and she takes the means to this end; she shoots as hard as she can with each joystick in virtue of the fact that these actions are productive of either a state of affairs in which target A is hit or a state of affairs in which target B is hit.[58] But, if the agent is performing an action with disjunctive structure, it doesn't look like we need

---

knowledge, and since at no point in the unfolding of the action does the agent know anything as specific as that she is hitting target A. See Thompson 2011. For my purposes, this is a side issue.

[55] See Bratman 1987, pgs. 113-16.

[56] See chapter nine of Bratman 1987.

[57] Bratman introduces endeavorings as part of a larger account of the relation between intentions and intentional actions called the Single Phenomenon View that he develops across chapters eight and nine of Bratman 1987. However, the attraction of the Single Phenomenon View is largely that it can make sense of cases that are structurally similar to the video game example. See Bratman 1987, pg. 137.

[58] The details of this case might give one pause. It might seem that the agent was performing some intentional actions in virtue of the fact that they were productive simply of hitting target A and some intentional actions in virtue of the fact that they were productive simply of hitting target B. The agent isn't performing any actions that are neutrally productive of either state of affairs coming about, and her shooting with joystick A doesn't seem to be sensitive to her continuing to be able to shoot with joystick B, and vice versa. But, I think we can chalk this up to the quirky nature of the thought experiment: if an agent's end is to either hit target A or hit target B and she wants to achieve this goal the most efficient way possible, the situation dictates that the agent behave just as she would if she were in the process of hitting both targets, because there are no available means that are neutrally productive of either state of affairs and nothing that the agent could do with one joystick affects what she does with the other joystick. And since the world ultimately "chooses" which disjunct is realized—not the agent in some future phase of the action, as in the cases I've discussed so far—and the agent doesn't care which disjunct is realized, she need not be careful not to let one of the two targets be

endeavorings at all. Since present-directed intentions have presently unfolding actions as objects, all we need is a present-directed intention to perform an action with disjunctive structure.[59]

Why, then, does Bratman think that he needs an additional mental state to make sense of the video game example? The move seems unmotivated, given the Anscombian account of structure. It does seem quite plausible, however, if Bratman holds the assumption that actions can only involve means taken simply to produce the state of affairs that the agent counts has having produced intentionally when the action is completed. If the agent's eventual hitting of target A can, in its unfolding over time, only involve means taken to hitting target A, the only plausible candidates for these means in this case are the agent's shooting with joystick A. But, if these actions are explained by one mental state, then it looks like we are going to need two different mental states, for the reason expressed in Bratman's argument above: the agent is also performing means that are productive of a state of affairs in which target B is hit, and these call for the same sort of explanation. The mental state cannot be an intention, however, since this would violate strong consistency. To solve this problem, we could posit another mental state, the rationality of which is not governed by strong consistency, which is what Bratman does.

The way Bratman initially sets up both the Simple View and the video game example further suggests that he holds this assumption. Given the Anscombian view of structure, the Simple View, as Bratman states it, is crucially ambiguous. When an agent successfully ϕs intentionally, ϕ stands for a completed instance of an action in the perfective aspect. But, when an agent has a present-directed intention to ϕ, ϕ stands for an action that is currently unfolding in the imperfective aspect. On the Anscombian account of structure, the action that is currently unfolding might not only involve means taken simply to produce the state of affairs that is realized when the action is completed. Given that an intention has a presently unfolding action as its object, it looks like there are two different matches between intention and action that might obtain. An agent's intention might match the action that is currently unfolding in virtue of having an action with the structure of the action that is currently unfolding as its object. Or, an agent's intention might—in some sense—match the action in the perfective by having as its object an action with a structure that involves means taken simply to hitting target A, where that is the structure of the action that is currently unfolding. Which match is required by the Simple View? Bratman doesn't say. However, the need to make this distinction would not arise if one assumed that

---

hit until some future point in time is reached. Thus, the agent's behavior looks to an observer deceptively like the behavior of someone who is in the process of hitting target A and hitting target B. But consider the following counterfactual: halfway through the game, the agent notices that one of the joysticks is non-responsive, and concludes that contrary to initial assessments, the most effective way of hitting either of the targets would be to start taking means simply to hitting target A, so she stops shooting with joystick B. If, as Bratman thinks, this agent was taking means independently to hit target A and to hit target B, explained by parallel mental states, then what explains why the agent stops taking the means to hit target B? The only answer that Bratman has, I think, is to acknowledge that the agent was not taking means simply to hitting target B in the first place, but to hitting either of the two targets, which is just to say the agent was carrying out an action with disjunctive structure.

[59] At least Bratman talks as though the object of a present-directed intention is an agent's presently unfolding action. See chapter 8 in Bratman 1987, where he generally talks about intentions to hit target A, where hitting target A is the action which the agent is currently carrying out. But, in general, what the object of an intention is supposed to be is not made explicitly clear in most of the writers that I will consider. Philosophers frequently use the common infinitival construction—X intends to ϕ—which suggests that a present-directed intention has a presently unfolding action as its object, but without ever making their commitments explicit. In a forthcoming paper arguing that actions are *not* the objects of intentions, Luca Ferrero notes that this is a common assumption and that philosophers are often less than careful when discussing the objects of executive attitudes. See Ferrero 2012b, pg. 6. For the ease of exposition in this paper, I will assume that all of the philosophers I discuss hold that the object of a present-directed intention is a presently unfolding action. I don't think that assuming this begs any questions, though, because it seems to me that regardless of what we take the object of the intention to be, it must be able to account for the intentionality of the action that is currently unfolding, and on the Anscombian account of structure, currently unfolding actions may have disjunctive structures.

actions in their unfolding could only involve means taken simply to produce the state of affairs which is realized when the action is completed.

Further, on neither reading of the Simple View offered in the previous paragraph does the video game example show that the Simple View is false in virtue of violating strong consistency, unless the Simple View is supplemented with the tacit assumption I've been discussing. On the first reading, the video game example does not show that the Simple View is false at all. The video game example shows that there are some actions in which the agent was never taking means simply to the end of hitting target A, but this doesn't mean that, throughout the unfolding of the action, the agent didn't have an intention to be carrying out an action with the structure of the action that she was in fact at that time carrying out. All that is shown is that there are actions which have disjunctive structures throughout the entirety of their unfolding in the progressive. On the second reading, the video game example does show that the Simple View is false, because the video game example shows that there are actions that have disjunctive structures at every point in their unfolding. But, on this reading the Simple View is false not because it dictates that an agent has intentions that violate strong consistency, but because some actions don't have the sort of structure that the Simple View requires.[60]

Given the assumption that, in its unfolding over time, the intentional action of hitting target A can only involve means taken simply to the end of hitting target A, however, the Simple View requires that the agent have a present-directed intention to carry out an action with that structure. But, given considerations of symmetry that Bratman cites—that the agent is behaving identically with respect to the goal of hitting target B—it looks like we should also attribute to the agent a present-directed intention to carry out an action whose structure involves means taken simply to the end of hitting target B. Since the agent knows that she cannot successfully complete both of these actions, the Simple View requires conflicting intentions.[61]

In summary, it looks like Bratman's argument requires the problematic assumption in order to be plausible. Given the Anscombian account of structure, however, this assumption is false.

3.2. Hugh McCann's Softening of Rational Requirements

Hugh McCann offers a defense of the Simple View against Bratman's argument. McCann argues that the norm of strong consistency governing the rationality of intentions is potentially defeasible and that there is no better evidence for thinking that there are in fact exceptions to this norm than the video

[60] One might think that the end that the agent is taking the means to at the moment at which an action is completed fixes the correct description of the state of affairs that the agent counts as having intentionally produced, and so in the video game example, the agent has only intentionally realized a state of affairs in which either target A or B is hit. I think Anscombians think this; see Thompson 2011. If this were the case, the video game example would not be a counterexample to the assumption I'm discussing. But, some of Bratman's key reasons for thinking that the video game example is an instance of hitting target A intentionally—e.g. that observational knowledge of target A being hit causes the agent to stop acting—are independent of considerations of what means the agent was taking at the moment at which the action was completed (117-19). So, if he were to recognize disjunctive structure, I think he would conclude that the video game example is an instance of hitting target A intentionally with entirely disjunctive structure.

[61] Also, it is worth noting that Bratman's discussion of why the video game example counts as an instance of hitting target A intentionally seems oblivious to the possibility that an action might not only be made up of means taken simply to produce the state of affairs that is intentionally realized. Bratman juxtaposes an agent's behavior being guided specifically by target A in the video game example with an example in which an agent's behavior is guided by a combination of two targets that are too close to distinguish, without seeming to realize that the agent's behavior being guided specifically by target A is consistent both with instances in which agents are simply taking the means to hit target A and instances in which they are trying to hit either target A or B. See pgs. 117-19.

game example itself; the example shows that it is sometimes rational to form inconsistent intentions.[62] So, Bratman's introduction of endeavorings is insufficiently motivated, because one need not posit endeavorings if it is acceptable to form inconsistent intentions.

But, the specific details of McCann's critique are not important here. What is important is that even though he avoids Bratman's bloated mental ontology, he still agrees with Bratman that the Simple View requires that in the video game example the agent have inconsistent intentions, for the reasons of symmetry that Bratman cites. However, as I argued in the previous section, the Simple View only requires that an agent have conflicting intentions given Bratman's tacit assumption. McCann's loosening of the rational requirements on intentions is only required, then, if one assumes that actions in their unfolding can only involve means taken simply to produce the state of affairs which the agent ultimately counts as having produced intentionally.[63]

3.3. Richard Holton's Partial Intentions

Finally, a related oversight occurs in an argument Richard Holton makes for the introduction of partial intentions.[64] Holton considers the following scenario: you want to remove a tree that has fallen down and is currently blocking your driveway, trapping your car. You conclude that there are four plausible ways to move the tree: lever it with a crowbar, cut it into pieces with a chainsaw, tie a rope to it and pull it with your car, or pay a bunch of money to the local tree company to move it for you. You start acting on all of these possibilities: you'd prefer not to have to pay, but you call the tree company and make an appointment just in case. Then, since you aren't sure that you will succeed at any of the three ways of removing the tree, you gather up your chainsaw, some rope, and a crowbar and head to the tree.

What intention do you have in this case? Holton wants to use this case to motivate the introduction of *partial intention*, a practical attitude that is supposed to be analogous to the widely accepted partial belief. In this scenario, Holton thinks that you have an all-out intention—that is the planning attitude we are familiar with from Bratman's work and the subsequent traditionalist literature—to remove the tree. But, in addition, you have four partial intentions: to lever it with a crowbar, to cut it into pieces with a chainsaw, to tie a rope to it and pull it with your car, and to have the tree company move it.[65]

---

[62] See McCann 1998.

[63] One might worry that the problems Bratman is concerned with are simply being pushed back a step: if an agent can form the intention to carry out an action which currently has disjunctive structure, it looks like she can form an intention to perform an action with a disjunctive structure of the sort that consistency norms on intention were supposed to rule out as irrational. But, as both McCann and Luca Ferrero have pointed out, the intuition behind consistency norms is tied to self-defeat: it is irrational for an agent to set out on mutually incompatible projects, because her efforts will be frustrated. (See McCann 1998 and Ferrero 2010, pgs. 12-21.) This point should carry over to the structure of actions. It is usually self-defeating to take the means to either of two different states of affairs all the way up to the point at which it would be impossible not to realize one of the states of affairs, and so agents who intend to perform actions with disjunctive structures past a certain point are irrational in virtue of pursuing a self-defeating course of action. But, there is nothing in itself irrational about taking means in virtue of the fact that they are productive of either of two state of affairs, and in the video game example, it is rational to perform an action with disjunctive structure throughout, since it is not self-defeating.

[64] See chapter two of Holton 2009.

[65] Holton 2009, pgs. 34-7.

Like endeavorings, partial intentions are what are present when agents are aiming at co-impossible ends.[66] But, like the video game example, Holton's example seems to be an action with disjunctive structure. Why not think that in this situation, the agent has an all-out intention to perform an action with disjunctive structure that involves taking the means to *either* levering the tree with a crowbar, *or* cutting it into pieces with a chainsaw, *or*... etc.? Holton considers the possibility that a single present-directed intention—what he calls a disjunctive intention—might be able to explain what the agent is doing, but rejects this option. It will be useful to quote his reasons for doing so in full:

> Of course we could say that [you have a disjunctive intention]; but to say that would be to lose explanatory force. For we need to break compound intentions down into their elements if we are to understand quite what explains what. Consider a parallel example with ordinary all-out intentions. Here presumably conjunction is permissible: if I intend to hear a concert and intend to buy some whisky, then I intend to hear a concert and buy some whisky. But we would not want to be constrained to use only the conjunctive sentence. It is my intention to hear the concert that explains why I buy a ticket; it is my intention to buy some whiskey that explains why I divert to the off-license. It is only if we break down the intention into consistent atoms that these explanations become available. The same is true when we try to give all-out disjunctive surrogates for partial intentions. It is my partial intention to get the tree company to move the tree that causes me to phone them; if we are limited just to all-out disjunctive intentions, we can give no explanation of this. (38)

Holton's reasoning in this paragraph is fairly opaque, but it amounts, I think, to something like the following:

1) The intentionality of an agent's ϕ-ing is explained just in case an agent has an intention the object of which is only that action: ϕ-ing.

2) Intentions with logically complex objects have more than one action as their object.

3) Therefore intentions with logically complex objects do not explain the intentionality of an agent's ϕ-ing.[67]

I am not exactly sure why Holton thinks that the first premise is true. The paragraph quoted above seems to involve the assertion of the premise coupled with elaboration that is supposed to appeal to the reader's intuitions: when an agent buys a bottle of whisky, the intentionality of this action is explained by an intention that has just that action as its object.

However, I think that we can grant premise one to Holton for the sake of argument. Premise two is another story. This claim seems to be supported in the paragraph above with appeal to the intuition that when an agent would express the intention to ϕ and ψ, ϕ-ing and ψ-ing are two different actions, and the agent is not expressing the intention to do something over and above those two actions. But, while this is

[66] Holton is less concerned with worries about the Simple View and more concerned with making sense of the intuition that we may be only "partially" committed to courses of action, but for my purposes this doesn't make a difference. See chapter two of Holton 2009.

[67] Holton's language suggests he might think that the object of an intention is not an action, but a state of affairs. But, see footnote thirty-two.

good evidence that in the particular example, when the agent would express a conjunctive intention, the object of that intention was two actions, rather than a single, conjunctive action, it is not clear why this is good evidence for the general claim that when an agent would express a present-directed intention with a logically complex object, the object involves multiple actions, rather than a single action with logically complex structure. Surely, this requires additional evidence, and Holton has failed to make his case that partial intentions are necessary.

However, Holton's argument seems plausible if he intends the example in the quoted paragraph not as a case from which to generalize about all logically complex intentions, but rather an appeal to his readers' intuitions in favor of the assumption we've been discussing. The conjunctive intention doesn't simply happen to have multiple actions as objects in this case, but it *must*, since all actions are like those in the example: involving only means taken to the state of affairs that is eventually realized. The intentions that have these sorts of actions as their sole objects always have logically simply form; the intention to buy a bottle of whisky has as its sole object an action involving means taken simply to produce a state of affairs in which a bottle of whisky is purchased. So not only the conjunctive intention, but all logically complex intentions, must have multiple actions as their objects. They therefore lack explanatory value. However, if the Anscombian account of structure is correct, Holton's intuition pump is faulty. Actions can have logically complex structure, and logically complex intentions have an explanatory role to play.

4. Conclusion

In this paper, I argued that on a Anscombian account of structure, actions may have logically complex structures, and then I surveyed four different arguments made by traditionalists that only work if we assume this to be false, an assumption these philosophers likely make for lack of considering means-end structure at all. These are only the most apparent examples: arguments made by major action theorists that focus on instances of action where, due to special circumstances, the disjunctive structure becomes much more apparent than it otherwise would be. But, I suspect that this assumption infects other arguments in ways that are not as readily apparent. In fact, given that traditionalists frequently talk about present-directed intentions to ϕ, where ϕ-ing is what the action will be an instance of in the perfective, even though the action may, at the present moment, not involve means taken to an end that is that specific yet, I suspect that similar problems may arise in many discussions of present-directed intentions.

I hope, then, that this paper has shown this much: no action theorist can ignore structure. The current dearth of interactions between traditionalists and Anscombians is somewhat understandable. Although they work on and disagree about many of the same issues, these camps work from some very different basic assumptions, and so many arguments by thinkers in one camp may seem to be missing the point to philosophers in the other camp. And so, perhaps many traditionalists take there to be little reason to engage with the Anscombians. However, as this paper has discussed, structure is—at least *prima facie*—a neutral desideratum for which all theories of action should account. If I am right, traditionalists need either to reject the Anscombian account of structure or—what I think is much more likely—seriously reconsider a number of their arguments.

Works Cited

Anscombe, G.E.M. 1979. "Under a Description." *Noûs* 13 (2): 219-233.

- - -. (1957) 2000. *Intention*. Cambridge, MA: Harvard University Press, 2000.

Bratman, Michael. 1987. *Intentions, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

- - -. 1999. "I Intend that We *J*." In *Faces of Intention: Selected Essays on Intention and Agency*. New York: Cambridge University Press.

- - -. 2012a. "Time, Rationality, and Self-Governance." *Philosophical Issues* 22 (1): 73-88.

- - -. 2012b. "The Fecundity of Planning Agency." Colloquium talk, University of Wisconsin – Madison, 12 October.

Chant, Sara Rachel. 2010. "Two Composition Questions in Action." In *New Waves in Metaphysics*, edited by Alan Hazlett, 27-53. New York: Macmillan.

Davidson, Donald. 1980. *Essays on Actions and Events*. New York: Oxford University Press.

Ferrero, Luca. 2009. "Conditional Intentions." *Noûs* 43 (9): 700-741.

- - -. 2010. "Disjunctive Intentions." Unpublished manuscript.

- - -. 2012a. "Diachronic Constraints of Practical Rationality." *Philosophical Issues* 22 (1): 144-164.

- - -. 2012b. "Must I Only Intend My Own Actions? Intentions and the Own Action Condition." Forthcoming in *Oxford Studies in Agency and Responsibility*.

Ford, Anton. 2011. "Action and Generality." In *Essays on Anscombe's Intention*, edited by Anton Ford, Jennifer Hornsby, and Frederick Stoutland, 76-104. Cambridge, MA: Harvard University Press.

Frey, Jennifer. 2012. "Practical Knowledge and the Good." Unpublished manuscript.

Galton, Antony. 1984. *The Logic of Aspect: An Axiomatic Approach*. New York: Oxford University Press.

Haddock, Adrian. 2011. "The Knowledge That a Man Has of His Intentional Actions." In *Essays on Anscombe's Intention*, edited by Anton Ford, Jennifer Hornsby, and Frederick Stoutland, 147-69. Cambridge, MA: Harvard University Press.

Holton, Richard. 2009. *Willing, Wanting, Waiting*. New York: Oxford University Press.

Kenny, Anthony. 1963. *Action, Emotion and Will*. London: Routledge & Kegan Paul.

McCann, Hugh. 1998. "Settled Objectives and Rational Constraints." In *The Works of Agency: On Human Action, Will, and Freedom*, 195-212. New York: Oxford University Press.

McDowell, John. 2010. "What is the Content of Intention in Action?" *Ratio*, 2010, 23.4: 415–432.

- - -. 2011. “Anscombe on Bodily Self-Knowledge.” In *Essays on Anscombe’s Intention*, edited by Anton Ford, Jennifer Hornsby, and Frederick Stoutland, 128-46. Cambridge, MA: Harvard University Press.

Millgram, Elijah. 2012. “Practical Reason and the Structure of Actions.” In *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/practical-reason-action/

Moran, Richard. 2004. “Practical Knowledge.” In *Agency and Action*, edited by Anthony O’Hear, 48-68. New York: Cambridge University Press.

Moran, Richard and Martin Stone. 2011. “Anscombe and Expression of Intention: An Exegesis.” In *Essays on Anscombe’s Intention*, edited by Anton Ford, Jennifer Hornsby, and Frederick Stoutland, 33-75. Cambridge, MA: Harvard University Press.

Mourelatos, Alexander. 1978. “Events, Processes, and States.” *Linguistics and Philosophy*, 2 (3): 415-34.

- - -. 1993. “Aristotle's Kinêsis/Energeia Distinction: A Marginal Note on Kathleen Gill's Paper.” *Canadian Journal of Philosophy* 23 (3): 385-88.

Setiya, Kieran. 2007. *Reasons Without Rationalism*. Princeton: Princeton University Press.

- - -. 2008. “Practical Knowledge.” *Ethics* 118 (3): 388-409.

- - -. 2009. “Practical Knowledge Revisited.” *Ethics* 120 (1): 128-37.

- - -. 2010. “Intention.” In *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/intention/

- - -. 2011. “Knowledge of Intention.” In *Essays on Anscombe’s Intention*, edited by Anton Ford, Jennifer Hornsby, and Frederick Stoutland 170-97.Cambridge, MA: Harvard University Press.

Speaks, Jeff. 2009. “Thompson on Aspect and the Primacy of Naïve Action Explanations.” Unpublished manuscript. http://nd.edu/~jspeaks/courses/2008-9/43503/index.htm

Stout, Rowland. 2005. *Action*. Ithaca, NY: McGill-Queen’s University Press.

Stoutland, Frederick. 2011. “Introduction: Anscombe’s *Intention* in Context.” In *Essays on Anscombe’s Intention*, edited by Anton Ford, Jennifer Hornsby, and Frederick Stoutland, 1-22. Cambridge, MA: Harvard University Press.

Tenenbaum, Sergio. 2010. “The Vice of Procrastination.” In *The Thief of Time*, edited by Chrisoula Andreou and Mark White, 130-50. New York: Oxford University Press.

- - -. 2012. “Reconsidering Intentions.” APA Symposium, “Choice Over Time,” Seattle.

Thompson, Michael. 2008. *Life and Action*. Cambridge, MA: Harvard University Press.

- - -. 2011. “Anscombe’s *Intention* and Practical Knowledge.” In *Essays on Anscombe’s Intention*, edited by Anton Ford, Jennifer Hornsby, and Frederick Stoutland, 198-210. Cambridge, MA: Harvard University Press.

Velleman, J. David. 1989. *Practical Reflection*. Chicago: University of Chicago Press.

Vendler, Zeno. 1967. *Linguistics* in *Philosophy*. Ithaca, NY: Cornell University Press.

Vogler, Candace. 2001. "Anscombe on Practical Inference." In *Varieties of Practical Reasoning*, edited by Elijah Millgram, 437-64. Cambridge, MA: The MIT Press.

- - -. 2002. *Reasonably Vicious*. Cambridge, MA: Harvard University Press.

- - -. 2009. "Nothing Added: §§19 and 20 of *Intention.*" Conference on Anscombe's *Intention*, University of Chicago, 25 April. https://mahimahi.uchicago.edu/admin/pcast/pcastpreview.php?podcastid=262&version=audio

Wilson, George. 1989. *The Intentionality of Human Action*. Palo Alto: Stanford University Press.

Wiseman, Rachael. 2017. *Routledge Guidebook to Anscombe's* Intention. New York: Routledge.

"Personal and Impersonal Good"
Nandi Theunissen

INTRODUCTION

One of the doctrines tightly associated with G. E. Moore is that there can be value in a world without valuers—actual valuers or possible ones. If Moore made a concession to the thought that value somehow essentially depends on possible appreciation by someone, it was that valuers add to the value of what is of value through their appreciation and enjoyment of it. A sunset seen and appreciated is better than one unseen and so on, but the sight that is not and cannot be seen is still of value. It must be said that unreconstituted Mooreans about value, as about most things, are by now exceedingly rare. Moore provoked a century of reflection on the nature of value, and whatever the differences between his interlocutors, they join in a common feeling that he left value in the dark. Of course there is more than one source of complaint. But a central complaint is that Moore failed to see that values are the kinds of thing that essentially matter to us, or to someone. To that extent he failed to explain what it is for something to be good or of value.

Giving expression to felt consensus is a fraught business; consensus in philosophy is often local and it is nearly always unstable. But I will venture something like broad agreement among value theorists, against Moore, on the following point: value necessarily depends on possible appreciation by someone. This is the thought, in a familiar example, that there would be no value in the Frick collection if all sentient life were destroyed (Nagel, 1986, 153). It is the thought that the works of art would be dead and worthless things since they are there, in some important sense, for us (Wolf, 2010, 56). In terminology due to Joseph Raz, the point is made by saying that value is "personal'' rather than "impersonal'' (Raz, 2001, 274). All values are personal in the sense that they depend on possible appreciation by someone, and no values are impersonal in the sense that their value is wholly independent of possible appreciation by anyone. How uncontentious a claim this is in part depends on who is being counted as a someone. I will not try to settle that question here, and for my purposes the net can be cast quite wide.

To accept that values necessarily depend on the possibility of subjects who can appreciate them is not to dispense with Moorean ideas about value, however. Indeed, what interests me in this article is that a new way of being a Moorean has made itself felt among those who accept the thesis that all values are personal. The position comes into view as follows. If something is of value then it must be possible for someone to appreciate it. But the value is or can be appreciated only if and because it is of value in itself. Or better, allowing that value can be instrumental and non-instrumental, whoever can or does appreciate something that is non-instrumentally valuable does so appropriately because and insofar as the thing is good simpliciter [this risks confusion]. My larger aim in what follows is to assess this style of proposal. Does it succeed in making goodness or value (terms I am using interchangeably) any less dark than Moore himself made it? That is, does it succeed in bringing out for us, as I believe it should, the sense in which values essentially matter to us, or to someone?

GOOD AND GOOD FOR

Let me make my question more precise. I said that I am responding to a new way of being a Moorean—in Nouveau Mooreanism. I call it new not because the relevant proposals are hot off the press (though some of them are), but because the position can be articulated in light of a schema or set of terms that have come into focus in recent discussions. The relevant terms are "good" and "good for." Start with the premise accepted by Nouveau Mooreans that all values are personal. If something is of value then it must be capable of being appreciated by someone. The personal character of value is taken to entail that whatever is of value is or can be good for someone. The entailment depends on the supposition that appreciating something of

value is good for the one who appreciates it. This may be put by saying that engaging with objects or activities of value is part of the good of the one who engages, or part of what enriches their life. So, if something is of value then, necessarily, it is or can be good for someone (beneficial for them, advantageous, salubrious, etc). Values can be instrumentally or non-instrumentally good for valuers, and when something is non-instrumentally good for a valuer it is directly good for them, by itself good for them, or good for them for its own sake. The key Nouveau Moorean claim is that when something is non-instrumentally good for a valuer, it is so because it is good in itself, or simpliciter. Goodness simpliciter is a form of value that is not a function of being good for someone. For Nouveau Mooreans it necessarily features in an explanation of why something is good for someone when it is. So, there are two properties: the properties of being good simpliciter, and the property of being good for someone. For Nouveau Mooreans, when something is good for someone, its being good simpliciter explains why it is so. To that extent, good simpliciter is more fundamental than and explanatorily prior to good for someone. I will call this position G, and those who defend it G theorists.

G theorists occupy an interesting intermediate position. On the one side are full-blown Mooreans who reject the thesis that all values are personal and reject the would-be entailment that whatever is of value is or can be good for someone. Indeed, classic Mooreans want nothing to do with the notion of the good for someone. On the other side are theorists who take the view that what it is for something to be good or of value just is for it to be (or be capable of being) good for someone. Call this position GF, and those who defend it GF theorists. For GF theorists it is false that there are two properties, the property of being good and the property of being good for someone. Rather, good is good for someone. The G theorist occupies a position between classic Mooreanism and GF theory. As with all intermediate styles of proposal—which are the more interesting for being intermediate—G is vulnerable to attack from both sides. Some, but relatively few, will attack G by defending Moore-unreconstituted. The more lively angle, I suspect, will be made by those who attack G by defending GF. I will focus on this side of things.

GF theorists are most friendly towards the supposition that all values are personal. They certainly accept the entailment that whatever is of value is or can be good for someone. But they opt for a more thoroughgoing rendering of this idea. GF theorists propose that whatever is of value is so because it is or can be good for someone, so that "good" is short for "good for someone." That to be good is to be good for someone, or equivalently, that to be good is to be beneficial, is an ancient Greek contention about value. It is prominently associated with Socrates, and it has enjoyed a resurgence in contemporary discussions (Kraut, 2007, 2011; Vogt, 2017. Cf. the symposium on Kraut including the contribution by Stroud).

The dialectical situation between G and GF theorists may be put as follows. GF theorists make a claim about what goodness or value *is*: to be good is to be good for someone. G theorists reject the proposed identification. From the G theorist's point of view, GF theorists are fleeing too far from Moore. G theorists raise a pointed question: Is it (A) good because it is good for someone, or (B) is it good for someone because it is good? G theorists defend B: it is good for someone because it is good. I examine the G theorist's arguments for B in what follows. I make the case that G theorists bring to light some crucial refinements to a conception of the good as good for someone. But ultimately I contend that the G theorist's arguments are insufficient to secure B. If the question is whether it is true that there must be two distinctly different kinds of value, and that one kind depends on the other, then the G theorist does not secure this conclusion. It is false that when something is good for someone it must be so because it is good. I conclude by locating the deep point of disagreement between G and GF theorists. For GF theorists, value is crucially and essentially affective. For G theorists, value affects us as a mere symptom of being good. Without settling the question of which is the better theory of value, I suggest that the G theorist is under pressure from the GF theorist to explain the claim of values on our cognitive and practical attention. If the suggestion stands, Nouveau Mooreans must do more to make a real advance over Moore.

AN ARGUMENT FOR G

G theorists invite a comparison with Moore insofar as they invoke a notion good simpliciter. But they also invite comparison with Moore insofar as they take an interest in aesthetic value, and one is drawn to recall that Moore was a figure of inspiration to the Bloomsbury group whose most famous member was Virginia Woolf. Like Moore, present-day G theorists take the explanation of works of art, or higher culture, including works of fiction and philosophy, as a central evaluative case. Philosophy and art (as activities and products) would have no value in the absence of subjects who had the capacity to appreciate them (Wolf, 2010, 56). Engaging with works of art can be of value to people, and a life which includes that engagement can thereby be enriched (Raz, 1986, 201). But a work of art can be of value to someone only if and because it is of value in itself. As Joseph Raz writes: *"If something is intrinsically good for me it is so because it is good [in itself]—'it would be good for you to read this novel. It is really excellent'—and it is that very quality which make it good for others too. It would be good for you to read the book for the same reason it is good for me, i.e. because it is an excellent book*" (Raz, 2007, 275). And in Susan Wolf's words: *"what is good about [works of art or philosophy] cannot be explained in terms of its benefits to us. The order of explanation, I have argued, must go the other way around. It is only because and insofar as there is something good about [art and] philosophy [...] that [it] can also be good for us in noninstrumental ways"* (Wolf, 2010, 61). These are statements of G.

G is a metaphysical thesis to the effect that goodness is more fundamental than and explanatory of goodness for someone. How is it to be shown that the metaphysical thesis is true? G theorists support the thesis by appeal to observations about how we speak of and experience the forms of value that are in question. A claim about the structure of value is justified by a claim about what our practice with value is like. G theorists tend to take descriptions of our practice with value to provide direct support for the metaphysical thesis, and there may be a danger of having a description of our practices be just another expression of our metaphysical convictions. It will be important to attend closely to what our experience with these forms of value is like (cf. Barry Stroud, 2011, 10-13). But I am also less sure that we can read metaphysical theses directly from claims about our practice with value. The phenomena may be compatible with more than one style of proposal, or it may be better explained by a proposal that does not occur to us at first.

How, according to G theorists, do we talk about and experience the forms of value that are in question, centrally, works of art? G theorists draw attention to the primacy of goodness in our engagement with value, and their observation has first and second-person analogues:

First-person: A work of art's goodness is our own rationale for engaging with and thereby of benefiting from it
Second-person: A work of art's goodness is our reason for thinking it stands to be good for someone else.

As they stand, I submit that these claims about our practices will meet with some resistance. G theorists have us unhesitatingly form judgments about and make pronouncements on works as being good or excellent. But it will be readily pointed out that we are frequently more circumspect. We qualify our judgments by setting parameters to our own taste—to what we ourselves find appealing. Hume, who is always worth reading on the character of our practices—whatever we make of his theory of those practices—says we choose our favorite author or artist as we do our friend: from a conformity of sensibility and disposition (Hume, 1985, 242). And when we enjoin a work to others, we frequently appeal to considerations that have to do with its suitability for their interests or experience at the moment. These reservations are important. And I will come back to them. But first I will try to draw out a truth in the G theorist's observations.

The relevant thought comes out in Thomas Nagel's discussion of aesthetic value. Nagel is prepared to allow that the value of most things appears to be a function of their goodness for us. But he wonders whether

there are forms of value for which this does not appear to be the whole story. His discussion is intentionally aporetic, but contains the following suggestion:

Most of the apparent reasons that initially present themselves to us are intimately connected with interests and desires, our own or those of others, and often with experiential satisfaction. But it seems that some of these interests give evidence that their objects have an intrinsic value which is not a function of [...] their value for anyone. I don't know how to establish whether there are any such values, but the objectifying tendency produces a strong impulse to believe that there are—especially in aesthetics, where the object of interest is external and the interest seems perpetually capable of criticism in light of further attention to the object (Nagel, 1986, 153).

The thought is that works of art are external to us: they are objects of outer rather than inner sense. We take an interest in these works, we respond to them, but we can ask whether that interest is merited, and whether our responses are apt. We make these critical assessments by attending to the works themselves. Our interest is conditioned by the work. So the thought is that the value must lie in the work itself.

Now some will resist the idea that our aesthetic practices have this normative character. Others will allow that they do, but they will hold our practices in contempt—they will see them as systematically in error. I do not find either of these positions plausible. I think aesthetic judgments can be better or worse, and that our interests in a work can be more or less merited, our responses more or less apt.
Allowing this, I think Nagel's observation at least brings to light the limitations of a simple-minded GF proposal according to which for something to be of value is (and is no more than) for it to please, or interest, or in some other way enrich a valuer (where these are understood to be specific dimensions of benefit). On that sort of view, to engage appropriately with a work of art, one looks past its features and to how the work is affecting one. This loses the sense in which aesthetic appreciation and conversation is object-directed or -beholden.

Let me say more about how our attention to aesthetic objects is object directed or beholden. You might think that in our more robust forms of engagement, articulating our aesthetic judgment is an important aim. We want to understand our grounds for finding a work worthwhile and to refine them. We sometimes do this in conversation with others, or by reading what others have to say. We investigate their reasons for liking a work, or not, and these we may challenge, or take up, or develop differently. If that is right, to be adequate to the character of our responses to works of art, the G theorist should say that we make citations of goodness, but more than that, we conceive of goodness as a consequential property. We like it because we think it is good, and we think it is good because… Here we sometimes appeal to more specific evaluative properties: it iscomplex, striking, imaginative, insightful. Finding the right word, or being inventive with our terminology, are marks of right appreciation. In the simplest case, terms are predicated of formal features of the work that are genre dependent. Here we seek to be precise and comprehensive in our description, and this is an ability that takes practice. We need to learn how to notice distinguishing features, and to evaluate them in ways that are appropriate. Usually the relevant judgments are made through comparison with other works whose value may be estimated to be greater or lesser than the one that is in question. Plausibly we revisit our evaluations over time, recognizing that what strikes us as worthwhile at first is liable to change.

To be fully adequate to these issues one would need to engage with aesthetic theory, but my aims are more limited. Part of the point of the G theorist's observation is that when we engage with works of art we are drawn, and we draw others, to attend to the work itself so that our experience is world-or object-directed. I am adding that an important dimension of this practice is to name the object in ways that are appropriate to it and particular. This takes us beyond a citation of goodness and to an appreciation of the specific value

bearing features of a work. If this is on the right lines, then we need to amend the G theorist's observation about our practice with value. Raz has it that it would be good for you to read this book because it is excellent. But we can now imagine that the conversation continues: *Oh? What is so good about it? Well it involves the recollection of a childhood, but it is really a meditation on memory...* In that case—and I think this is uncontentious—the G theorist should refine her proposal in such a way that for something to be good or of value is for it to be good in itself in virtue of various features. Those features may be intrinsic (such and such style of composition) and extrinsic (that it was written by Proust). Then G theorists will hold that whatever is in this way of value can be good for people, and it will be good for them, when it is, in part because it is valuable in this way.

THE GF THEORIST REPLIES

So far I have made a friendly modification to the observation, and a friendly modification to G to take account of it. Now we can ask whether G is the only or the best way to make sense of the observation. The observation, to repeat, is that "*it would be good for you to read this novel because it is excellent, and it is excellent because…"* What might GF theorists say? It is open to a GF theorist to propose that *"the book is excellent"* is true not in case the book possesses the first-order property good, but in case it possesses some other practically relevant properties, for example, that it is formally innovative, and insightful about memory, etc. In that case, the observation is that *"it would be good for you to read this novel because it is formally innovative, etc."* What makes it true to say this? It is true just in case being formally innovative, etc., would make the book such that it can be good for the reader. Taken this way the datum does not secure an explanatory role for goodness simpliciter. The work has features that make it good for people to engage with, and it would not be good for people unless they pay attention to those features.

What emerges is a refined GF position according to which something is good not in virtue of possessing the property good, but in virtue of possessing some other practically relevant properties. To that extent being good is the second-order property of having properties that are practically relevant. So far the proposal is compatible with a "buck-passing" account according to which the normative burden is passed from goodness to some other set of natural or metaphysical properties.
Where the account would differ from a buck-passing account would be in requiring a further explanation of how those properties make the object practically relevant. On the present proposal, that further explanation is supplied by the relation of benefit. To be of value is to possess properties that are reason-giving because they can be good for an individual. According to this refined GF model, there are two dimensions to an account of the ground of value. The first looks to the features that make an object practically relevant. The second seeks an explanation of why those features make the object practically relevant in terms of the relation of benefit. But both components are relevant, indeed necessary, to an account of why something is good or of value. The account makes goodness for someone more fundamental than and explanatorily prior to goodness but it does so in such a way that it provides an ineliminable role for object-directed attention.

OF ESSENCE AND QUALITY

The discussion brings out that the G theorist's observations about our practice with value underdetermine the theory. Attention to those observations prompt refinements to simple versions of GF and G. But suitably refined versions of G and GF both stand to be adequate to the observations. So what is really at

stake between them? I will suggest that G and GF theorists have deeply different conceptions of value, and of the role of subjects and affection therein.

Go back to the G theorist's description of our practice with value. G theorists give voice to the thought that works of art call for a detached mode of engagement in which our attention is directed to the work itself and not to our enrichment. For example, Sarah Buss speaks of our appreciation of art as something whose value does not depend on our own needs and concerns (Buss, 2012, 354). Likewise, Susan Wolf writes of the quality and not the consequences that makes it appropriate to value works of art (Wolf, 2010, 51). In that case, to the extent we engage appropriately, we attend to the work and not to what the work is doing for us or how it is affecting us. Here G theorists may mean to capture the sense in which some interested modes of engagement obscure our ability to appreciate the qualities of a work. Hume, for example, writes memorably of how envy of someone else's work, or personal affiliation to its producer, are liable to throw our judgments off. But one may well wonder if G theorists are overstating the point. Plausibly, being moved or affected and in that way enriched by a work is part of what allows us to name the object in ways that are adequate to it. We are able to identify *this* as a relevant bearer of value because it is stirring or otherwise speaking to us.

So it may be urged that the G theorist's account of the phenomena is actually overly simple. To return to Raz's example, in recommending the novel to someone to read, we point to features of the work, but also features of the reader (that they enjoy this sort of thing, or have had an experience lately that the book would particularly illuminate, etc.). Or we point not just to the ways the book is likely to be good for this particular reader, but the ways the book is good for (enriching to) the expert reader—the critic. Likely, in making these sorts of recommendations we shift between talking about features of the book and talking about how these features make the book good for (stimulating for, enriching to, etc.) readers. That is, aesthetic appreciation is both object and *subject* directed.

G theorists find a central role for possible appreciation in their theory of value. But how do they understand the role of the subject in aesthetic experience exactly? For G theorists, the subject appreciates or recognizes the value of what is valuable. Here is Raz: *"Goods whose value is realized are not wasted goods. All this is a bit of a mouthful to say that paintings are there to be seen and appreciated, novels to be read, oranges to be eaten, mountains to be looked at or climbed. They are there for these things to happen to them in the sense that their value to others remains unrealized until someone [of value in himself] relates to them in the right way"* (VRA, 154). The role of a subject in aesthetic experience is to ensure that the value of the object is appreciated and not wasted. The subject is there to bear witness to the object's value which is self-standing. By bearing witness the subject is affected by the value (is benefited by it). This is not the essence of value but a symptom of it.

How do GF theorist's conceive of the role of the subject in aesthetic experience? For the GF theorist value is crucially affecting. Imagine you are reading Proust. The GF theorist conceives of the scene of aesthetic valuation like this. The activity of reading—an activity that involves comprehension, memory, attention, emotional responsiveness, and so on—is a process that involves a change in the state of the reader. Reading is a kind of thinking and the thinking gives rise to thoughts. The reader's imagination is set on fire. Her mind is pierced by a sudden realization and this changes the state of her understanding. In reading, the subject is being affected, and the value of the activity of reading lies in this: in the thinking, understanding and imagining. There are changes in the subject. There are various states of transformation. For the GF theorist, it is in these states that the value lies.

The discussion serves to locate the key point of difference between G and GF theorists. To formulate the difference, I borrow some vocabulary from the classic version of the puzzle the form of which has been in question. That is, I borrow some terms from Plato's *Euthyphro*. There, familiarly, Socrates is asking whether something is pious because it is loved by the gods or whether something is loved by the gods

because it is pious. Socrates finds it obvious that piety has priority over being loved by the gods. And to express his thoughts about the status of the property of being loved by someone he makes a comparison to a couch that is being carried by someone. From the fact that a couch is being carried by someone one learns something about couches, namely, that they are carriable or portable. But one does not thereby learn what a couch is. Likewise, from the fact that some action is being loved by someone, one learns something about the action, namely that it is lovable, but one doesn't learn what it is about the action that makes it so. Plato will put this by saying that the property of being god-loved, like the property of being carried by someone, is an affect property, where that is to say, it is a quality that something has that is in some way external to what it essentially is – to the what-it-is of its bearer.

To put the point in the terms that are at stake for me, I would say that for G theorists, the property of being beneficial for someone, the property of being good for someone, like the property of being loved by the gods, is a quality of good things that is external to the what-it-is of goodness—it is a property of all good things, but the property does not tell you what goodness essentially is. For the GF theorist, on the other hand, being benefitted by something is not an affect quality of being valuable; it is the what-it-is or essence of goodness. GF theorists are looking to bring the judgment of a work's value onto the same plane as the subject's being moved or drawn in or otherwise affected by the work. From their point of view, G theorists attend to this as an affect property when it is actually where the being of value is disclosed. For the GF theorist, goodness is an encounter between an object and a subject such that, because of the fit between them, the subject is enriched by the object.

To the extent that G theorists leave this out of an explanation of what goodness is, GF theorists will urge that there is something less satisfying about a G theorist's account of what value is. To invoke a remark from the early Korsgaard, "According to Moore the question why [something] has intrinsic value must not be raised: it just has the property of intrinsic value; there is no reason why it has that property (*Principia*, pp. 142-44). Yet it is because it has intrinsic value that we ought to make it an end in our actions. A thing's goodness becomes a property that we intuit and respond to in a way that seems curiously divorced from our natural interests" (Korsgaard, 1983, 194). My sense is that Nouveau Mooreans are vulnerable to the same kind of charge. They accept that value necessarily depends on the possibility of subjects to appreciate them. But arguably their way of conceiving of the role of subjects in appreciation falls short of explaining the claim of good things on our cognitive and practical attention.

CONCLUSION

I have been investigating a claim about the dependence of one form of value on another. There is more than one way to motivate a claim about value dependence. On the form of argument that has been in question here, when something is intrinsically good for a person, its being good simpliciter necessarily features in an explanation of why it is good for them, and the explanation of works of art is taken as a central case. I have argued that a theory of value must be sensitive to the object-directed character of appreciation and engagement. But the point falls short of showing that works of art must be good simpliciter. A suitably refined conception of the good as good for someone can capture object-directed attention. Moreover, there are grounds for thinking that explanations of the value of art in terms of their propensity to be good for us are more adequate to the sense in which values essentially matter to us.

REFERENCES

Sarah Buss, "The Value of Humanity," *Journal of Philosophy* CIX (2012): 341-377.

Christine Korsgaard, “Two Distinctions in Goodness, *The Philosophical Review* 92.2 (1983): 169-95.
Richard Kraut, *What Is Good and Why?,* Cambridge MA: Harvard University Press, 2007.
Richard Kraut, *Against Absolute Goodness,* Oxford: Oxford University Press, 2011.
David Hume, “Of the Standard of Taste,” *Essays: Moral, Political and Literary,* Eugene F. Miller ed. Indianapolis: Liberty Classics, 226-249.
Thomas Nagel, *The View From Nowhere,* Oxford and New York: Oxford University Press, 1986
Joseph Raz, “The Role of Well-Being,” *Philosophical Perspectives* 18 (2004): 269-294.
Joseph Raz, *the Morality of Freedom* Oxford: Clarendon Press, 1986.
Sarah Stroud, “‘Good For’ *Supra* ‘Good,’” *Philosophy and Phenomenological Research* 87. 2 (2013): 459-466.
Barry Stroud*, Engagement and Metaphysical Dissatisfaction,* Oxford: Oxford University Press, 2011.
Katja Vogt, *Desiring the Good,* Oxford and New York: Oxford University Press, 2017.
Susan Wolf, “Good For Nothing,” *Proceedings and Address of the American Philosophical Association* 85.2 (2010): 47–64.

MORAL LIABILITY IN WAR AND SELF-DEFENSE: EXTENDING THE JUST CAUSE ARGUMENT

KATHERINE SWEET

INTRODUCTION

In this paper I address a concern over how we ought to understand moral liability and its relation to the actions one may permissibly take during war and self-defense. First, I describe Jeff McMahan's account of liability to attack during war, which he thinks also applies to self-defense cases; I explain the traditional account of liability during war and juxtapose it with McMahan's alternative account. I then explain a condition that seems to be a requirement for liability, what I call the effectiveness condition, and I briefly defend a reading of McMahan on which this is a constitutive condition of a person's liability.

I then describe a self-defense case that McMahan uses, in which a sheriff would be liable to attack even though he poses no direct violent threat to the person about to be killed. I explain the effectiveness condition as it relates to the case of the sheriff, and I defend a reading of McMahan on which the effectiveness condition in war refers to the probability of helping to achieve (or not harming the achievement of) the broadly viewed *just cause*. I describe a second version of the sheriff case in which killing the sheriff would *not* save the mayor, and I describe what I take to be a common intuition in this case. I then provide a way for McMahan to explain this intuition, namely by incorporating the notion of a just cause into cases of self-defense. I conclude by arguing that the notion of a just cause in war can easily be extended to cases of typical self-defense, which might affect our intuitions about McMahan's argument.

THE STANDARD VIEW AND MCMAHAN'S ALTERNATIVE

In his book, *Killing in War*, Jeff McMahan defends a view of just war theory on which the moral liability to defensive killing or attack is determined not by the collective status of the agents involved, but rather by the individual actions taken by the agents.[68] An agent is morally liable to intentional defensive attack during war only if it would be permissible for a certain agent to violently attack him in certain conditions for certain reasons, where a violent attack can include killing.[69] On the traditional view of just war theory, a combatant is morally liable to intentional defensive attack if and only if he has retained combatant status at the time that he poses a threat to another during war. So although one side may be just and the other unjust, because all those fighting on the front lines are considered combatants, they are all equally liable to direct attack by the enemy.[70] This McMahan calls the principle of the *moral equality of combatants*.

McMahan argues that the traditional view is grounded in collective association with a group, rather than based in the actions of individuals. No matter what a civilian does to promote a war, he is not liable to intentional attack because of his status as a civilian. He may change his status and thereby

[68] See McMahan (2009), pp. 208 and 212. From here on out all references will be to McMahan (2009).
[69] See p. 4 and 10.
[70] See p. 4.

become liable as a combatant, but he still was not liable when he was a civilian. A person's moral liability is not the only thing that would make it permissible to attack him, for instance if a lesser evil argument permits a combatant to kill innocent civilians. However, the equal moral liability of all combatants is what explains most of the morally permitted acts that occur during war. For on this account, a normal soldier does not do wrong when he follows orders and begins shooting at the enemy, regardless of whether his government started an unjust war.

On the same token, he is liable to attack regardless of whether he is on the just or unjust side, and regardless of his moral permission to shoot the enemy. Thus, the soldier is not liable for the *jus ad bellum* violation that his government initiated by starting an unjust war. The soldier is only liable for violating *jus in bello* requirements, such as committing international war crimes.

McMahan rejects this account in favor of a view on which the actions of the individual are directly relevant to determining his liability to defensive attack. McMahan argues that self-defense cases provide the basis for our understanding of liability during war; cases of killing in war are just a special kind of self-defense. But in cases of self-defense, we certainly would not consider a would-be murderer and a police officer as having the same moral status, as the police officer attempts to shoot the murderer to prevent him from killing an innocent third party. The murderer has no permission to defend himself against the officer, since the officer has permission and justification for killing him.[71] Thus, the moral equality principle does not hold for combatants in a self-defense case, and therefore does not hold in cases of war (since on McMahan's account, war is a special kind of self-defense case).

McMahan defends an alternative account of liability during war, on which the moral liability to defensive attack that a person may have depends directly on (1) the actions he takes, (2) his moral responsibility with regard to said actions, and (3) the relation those actions have on the just combatants on the other side. For McMahan, a person is morally liable to defensive attack just in case he is morally responsible for an objectively unjustified threat of harm to another person.[72] This account of liability applies equally to the would-be murderer and the soldier in combat. If a combatant is fighting on the just side of a war, then he does not make himself liable to attack by wearing a uniform or fighting and killing those soldiers on the opposing side. Thus, the combatants during war do not share the same moral status in virtue of being a part of the same collective group.[73]

THE EFFECTIVENESS CONDITION: LIABILITY, PERMISSION, AND JUSTIFICATION

Though early on McMahan claims that the criterion for liability is the moral responsibility requirement described above, he later discusses the effectiveness of the defensive attack on preventing the harm as a requirement as well. Basically, an innocent person can only kill someone defensively if there is a certain potential for success, namely the success that the person's death will prevent the harm. A question arises from this account of liability: does the potential success of the defensive attack partly

[71] See p. 14.

[72] See p. 35.

[73] McMahan argues that such an account allows for the preservation of equality regarding legal liability, since we can hold such equality for pragmatic reasons. But the moral liability one has by being on the unjust side of a war cannot be determined by legal liability. See p. 189-192.

determine the agent's liability to such an attack, or is liability independent of the effectiveness of the attack?

If the latter, then perhaps the effectiveness requirement arises when considering an agent's justification (rather than mere permission) for such a defensive attack. Justified acts are those which a person has positive moral reason to do, whereas merely permitted acts are those which an agent would not wrong anyone by doing but does not have a moral reason to do. Perhaps, in other words, (1) a person is liable if and only if he is *morally responsible* for an unjustified threat of harm, (2) another is *permitted* to intentionally kill him if he is liable, and (3) the other is *justified* in killing him just in case he is liable *and* the effectiveness condition holds. Though I consider this a plausible view, McMahan seems to reject it later in the book. I will explain why I interpret him as claiming that the effectiveness is a constitutive requirement of an agent's liability rather than an independent claim representing a distinction between permission and justification.

McMahan first explains potential success of defensive attack as directly related to the condition of proportionality. He says,

> In most cases, for an act that causes harm to be justified, it must be instrumental to the achievement of some valuable goal against which the harm can be weighed and assessed… The principal condition of a person's being liable to be harmed in the pursuit of the goal is that he or she be implicated in some way in the existence of the problem. If a person is implicated in the existence of a problem in such a way that harming him in a certain way in the course of solving the problem would not wrong him, then he is liable to that harm.[74]

This account of liability seems to be at odds with his later claim that liability is determined by one's moral responsibility for an unjustified threat. For he states that the criterion for liability to defensive attack is the moral responsibility of an unjustified threat.[75] But if this is the criterion, then where does the effectiveness requirement come in? Is it another criterion, and if so then how many are there? This description certainly favors the interpretation laid out above, where the effectiveness condition is irrelevant to a person's liability.

However, when McMahan later discusses liability and effectiveness in his final chapter, he makes several claims that seem to endorse the effectiveness condition as constitutive of a person's liability. First, he claims, "moral responsibility for an unjustified threat that one does not oneself pose is sufficient for liability to harm *as a means of protecting the person* wrongly threatened [emphasis added]."[76] He also argues a similar point when distinguishing between desert and liability:

[74] Ibid., p. 19.
[75] See p. 35.
[76] Ibid., p. 207.

> [L]iability is not like desert in being determined only by what one has done; one's liability to harm is also a function of the harms that others will suffer, and for which one will bear some responsibility, if one is not harmed.[77]

In these cases, McMahan seems to be following the self-defense notion of liability, which fits with the overall structure of his account, since he thinks that war is a special case of self-defense. On this view, self-defense can create a setting for liability to defensive killing partly because one's life is being threatened and one has a right to defend it in some way against the person who is morally responsible for the threat posed. This differs from cases of punishment, in which punitive liability is determined by desert, which does not regard threats to one's life or other harm as necessarily relevant to the punishment a person is permitted to receive. So on McMahan's account, liability during war is very similar to self-defense liability in that it contains an element of potential success at preventing the wrongful threat that would otherwise occur.

This interpretation can be drawn from a case McMahan presents in his final chapter, which I will call the Sheriff Case.[78] A powerful sheriff convinces a poor farmhand through irresistible duress to assassinate the mayor of the town. But the mayor finds out the details of the sheriff's plan. Before he has a chance to confront the mayor, he sees the farmhand hiding with a gun, looking at him. He also sees the sheriff, and he knows that if the sheriff is dead then the farmhand will be released from his feelings of obligation to kill him. He also knows that the farmhand is very limited in his responsibility for what is about to happen, since he is under duress, and that the sheriff is responsible to a very large degree for his potential death.

It seems as though the mayor can and perhaps should defend himself by killing the sheriff rather than the farmhand. This is explained by the fact that the liability of the sheriff is partly due to the mayor's ability to kill him *to prevent* his own death. He knows that the effectiveness of preventing his death by killing the sheriff is high since the farmhand was threatened into helping the sheriff and has no allegiance to him. The sheriff is liable to defensive killing in part because he is standing right there, and his death has a good chance of saving the mayor. His liability is also due in part to his greater level of moral responsibility for the unjustified threat posed against the mayor, even though the farmhand poses the direct threat to the mayor's life in that moment. The sheriff is responsible to a greater degree than the farmhand, and the effectiveness condition holds for both of them, so the mayor is permitted to kill the sheriff.

This requirement (the potential success at preventing the initial wrong) is meant to hold in conditions of war as well. However, one problem with this account of liability is that, during war, several conditions of effectiveness may hold. One is the effectiveness of winning the current conflict in which an agent is fighting. If killing an unjust combatant would increase the chances that the just combatants will get off the battlefield safely that night and sleep in peace, then this perhaps satisfies the effectiveness condition that McMahan seems to hold. But what if it won't? Is shooting an unjust combatant on the battlefield impermissible if one thinks that winning this particular battle is impossible for the just side?

Though this version of the condition seems plausible, what McMahan seems to claim is that the just combatant is permitted to attack to prevent the unjust side *overall* from achieving its unjust aim, and thereby to promote the just aim of his own side. For, regarding civilians who are at least partly responsible for the unjust war their country wages, McMahan argues that "they cannot be liable to attack

[77] Ibid., p. 227.

[78] For the full case, see pp. 205-208.

unless attacking them can make an effective contribution to the achievement of *a just cause* [emphasis added]."[79] This seems to allow for a group of just combatants to attack first in a specific battle, and also to attack regardless of the potential success of the specific battle, except as that success is related to the overall just aim.

But if this is so, then it seems as though the effectiveness element to liability has been broadened so much that, although it is a condition of an agent's liability, it is so broad as to make virtually anyone liable so long as they satisfy the moral responsibility requirement. For instance, during war a just combatant may attack an unjust combatant knowing full well that he will achieve nothing from such an attack other than the fact that the unjust side will have *one* less combatant in their ranks. His life may not be directly threatened, and none of his comrades' lives may be threatened. In fact, McMahan thinks that it is more plausible that the effectiveness condition holds in cases in which the just cause is at stake rather than merely the individual lives of the just combatants.[80]

EXTENDING THE CASE: A RETURN TO SELF-DEFENSE

McMahan claims that the degree of moral responsibility an unjust combatant has, so long as it is nonzero, is nearly irrelevant when determining an unjust combatant's liability, so it is plausible to think that the effectiveness condition holds in this broad sense as potential success of the just cause.[81] But if unjust combatants with very limited responsibility and a low probability of affecting the outcome of the just cause can be liable in war, then perhaps an analogous point can be made in cases of self-defense.

Consider the Sheriff Case again. Let's assume that the mayor does know the effectiveness of killing the sheriff in preventing his own death, and it is small. The farmhand thinks that the sheriff has an accomplice who will kill him if he does not follow through with the plan to assassinate the mayor. So killing the sheriff will not keep the farmhand from killing the mayor. Let's also stipulate that the mayor can only kill one of them before one or the other has a chance to kill him. Ought he to kill the sheriff or the farmhand? It seems as though, regardless of the effectiveness of such an attack or his knowledge of such effectiveness, the sheriff is still liable and the mayor is still permitted in killing him over the farmhand, given that the mayor knows the sheriff to be morally responsible to a much greater degree than the farmhand.

But what explains this intuition? Perhaps the case of self-defense here has an analogy to the general just cause described above. That is, perhaps killing the sheriff, even if the effectiveness condition is low or negligible, is permitted because the mayor, through his action, is promoting a just cause (defending his life over the life of the evil sheriff). Regardless of the effectiveness of helping to save his own life, the mayor seems to be permitted to kill the sheriff because to do so would be to defend his own innocent life over the guilty life of the sheriff, and also to defend the threatened farmhand over the life of the sheriff. He may also be defending a just cause because the future state of the town will be partly determined by who survives, and if he kills the farmhand then the sheriff will likely run the town and wield power for nefarious purposes. This just cause (potentially saving the town) also seems to have

[79] Ibid., p. 225.

[80] This is because the just combatants could preserve their lives by not participating in war. So the participation in war shows that the just cause is much greater than the lives of the individuals participating in the war. See pp. 195-198 for further discussion of this issue.

[81] See p. 197.

greater weight than saving the lives of any of the individuals involved in the threat, including the mayor himself. So the mayor seems to have permission to kill the sheriff for the same reason that the just combatant has permission to kill an unjust combatant in almost any circumstances.

Perhaps the intuition in this case is mistaken, and we would not consider the mayor permitted to kill the sheriff. However, it seems as though the notion of a broadly conceived just cause in war could extend easily to cases of self-defense, which would allow us to lower the requirements for liability to explain such cases. If we leave our account of effectiveness in war broad enough that it encompasses anything that helps the just cause, then it seems that the effectiveness condition in self-defense ought also to be broad enough to encompass not just the immediate threat to one's life but also potential threats that might occur if the person responsible is not attacked. If true, then cases of self-defense can become broadly cases of defense of what is right over what is wrong, so long as the person who is morally responsible poses a *potential* threat to the immediate victim and to others.

McMahan's account of liability to attack during war seems to include the effectiveness condition, but the effectiveness condition is itself ambiguous. If it can be sufficiently broadened to encompass the general promotion of the just cause in war, then it seems that it could also be broadened to include self-defense cases in which the potential success constitutes thwarting a larger evil plan that would affect a large group, rather than the potential success of merely saving one's own life.

This might alter our intuitions regarding McMahan's alternative account of moral liability. For McMahan wants to argue that intuitively we should think of killing during war as special cases of self-defense killing. It is this claim that ultimately grounds the arguments for the *moral equality of combatants thesis* that has radically altered our ways of thinking about combatants. But if our intuitions in the case above prove to be against the view that a standard self-defense killing can be grounded in a broad "noble cause" rather than in a specific defense of one's own body, then we may have reasons for thinking that killing during war differs in an important way from self-defense. Or we may have reason for thinking that we should reject McMahan's arguments for a broad "just cause" defensive attack during war.

To consider the strength of this point, consider the sheriff example again. Suppose the farmhand shoots me (the mayor) fatally, but I have about 5 minutes before I will cease breathing. I now have the option to shoot the sheriff in order to prevent the town from falling into the hands of a dictator, who is morally responsible for an unjustified threat of harm. The effectiveness condition also holds, because my shooting him will make the whole town very happy and none will fill the void by taking over tyrannically in my absence. Perhaps what I do in shooting him seems justified, but is it really self-defense? McMahan, I've argued, would have to say yes, since it is satisfied by the two criteria I've described. But our intuitions may very well differ in this regard.

This possibility is enough to raise questions regarding McMahan's deep connection between self-defense cases and killing during war. For this reason, we ought to look further into this area to find if we really take token war deaths to be deaths from self-defense.

You're So Smug, I'll Bet You Don't Care This Paper Is About You

Grant Roseboom

Abstract: Unjustifiably expecting a higher form of regard from others than one deserves is a familiar vice; call it the "vanity-vice." Rousseau claims that the vanity-vice is uniquely morally dangerous. Using a Rousseau-inspired account of morally dangerous vices, I argue that his claim applies to only one form of the vanity-vice, which I call "entitled smugness." Entitled smugness is distinguished by the kind of deference it expects from others and the normative carelessness it exhibits.

1. Unjustifiably expecting a higher form of regard from others than one deserves is a familiar vice, which elicits such epithets as "arrogant," "vain," and "entitled."[82] Call this general trait the "vanity-vice." Vain people present a variety of problems. They are unpleasant to spend time with, they tend to be bad friends and colleagues, and they often discount the legitimate interests and concerns of others. But Rousseau (1979) goes further when he discusses the dangers of vanity in *Emile*. He thinks that it is a uniquely dangerous moral vice. He worries about this when, imagining that life has gone well for his fictional pupil, Emile, he writes,

> "Emile, in considering his rank in the human species and seeing himself so happily placed there, will be tempted to honor his reason for the work of [the tutor] and to attribute his happiness to his own merit. He will say to himself, "I am wise, and men are mad." In pitying them, he will despise them; in congratulating himself, he will esteem himself more, and in feeling himself to be happier than them, he will believe himself worthier to be so. *This is the error most to be feared, because it is the most difficult to destroy*." (p. 245, emphasis mine)

My aim here is to argue that Rousseau's claim that vanity is uniquely morally dangerous applies to only one form of this vice, which I call "entitled smugness." Entitled smugness is distinguished by the authority-recognizing deference it expects from others and the normative carelessness it exhibits, akin to Frankfurtian bullshit (Frankfurt 2005) and the epistemic vice of "insouciance." (Cassam 2018)

It is easy to ignore and/or exaggerate the dangers of vanity when we do not pay attention to the importantly different forms of the vanity-vice. For instance, as I discuss below, both Macalester Bell (2013) and Aaron James (2014) identify some of the central aspects and dangers of the vanity-vice (for Bell, the relevant vice is "superbia," and for James, it is being an "asshole"), but neither fully appreciates the importantly different kinds of vanity. As a result, they both fail to adequately account for the uniquely morally dangerous character of entitled smugness, and the comparably more benign character of the other forms of the vanity-vice.

[82] By "expecting," I do not primarily mean that the vain believe they deserve a higher form of regard than what they in fact deserve, although they do typically believe this. What I mean, as I explain in section 3, is that they take the norms governing the relevant forms of high regard to cast them as a worthy object of high regard and, accordingly, to call for others to hold them in high regard.

2. As I said above, there are many dangers and problems associated with the vanity-vice. What specifically is meant, then, in asking which form of this vice is most morally dangerous? To explain the relevant notion of moral danger, I construct a Rousseauian account that highlights two, interrelated dimensions along which vicious character traits might be morally dangerous: (1) they involve being unresponsive to standard forms of moral correction, and (2) they are incompatible with helping realize a republican ideal of egalitarian social relations.[83]

Let us begin with Rousseau's distinction between benign and inflamed amour-propre and the social conditions associated with each of these psychological states. Rousseau observes that, in being bound by social ties of any sort, we develop benign amour-propre – a concern to be taken seriously by others.[84] This is because being bound by social ties makes us worried about being overlooked by those on whom we depend, where "overlooked" means, roughly, not having one's voice (i.e., judgments, decisions, interests) taken into account. This benign form of amour-propre underlies a kind of healthy self-respect. But Rousseau argues that, in typical social conditions, benign amour-propre is not a stable frame of mind. Persons tend to be taken seriously just when they attract others' attention and consideration. Once we impress, charm, or flatter others in some way, our voice is taken into account; otherwise, our voice is ignored. This general pattern of interaction makes us anxious about being taken seriously by others. For, not only is it difficult to continually attract others' consideration, we also may find ourselves stacked up against those who are more talented, witty, smart, etc. We thus face a persistent, uncertain threat of losing our basis for being taken seriously by others, and our anxiety about this threat makes us desire to exercise more control over how others think and feel about us, which is inflamed amour-propre.[85] We desire to eliminate others' latitude to not take us seriously, and this amounts to a desire for what republican thinkers often call "domination."[86]

The social conditions associated with inflamed amour-propre, then, are conditions in which persons tend to be taken seriously only to the extent they show themselves to be admirable, lovable, or remarkable in various ways. These conditions produce widespread anxiety about being well-regarded by others that, in turn, creates a strong desire for domination. Elsewhere, I argue that to alleviate these social conditions, we need a social practice of basic equality.[87] (Rozeboom 2018) We need a practice by which all individuals who share and can help solve the problem of inflamed amour-propre are given the same basic standing to govern their lives and help structure their interactions with one another. (Ibid., pp. 154-7) What is important for our purposes is seeing that this egalitarian social practice requires individuals to accept norms of mutual deference. In accepting that all persons have the same basic standing to govern their lives, individuals constrain themselves from paternalistic interference in one another's affairs. And in accepting that all persons have the same basic standing to help structure their interactions with one another, individuals constrain themselves from unilaterally dictating the terms of their relations to others.

This suggests that, on a Rousseauian account, a central condition of morally desirable character traits is that they involve maintaining attitudes of regard toward oneself and others that are compatible with, and help constitute, the kind of mutual deference entailed by a social practice of equality. That is, how one is disposed to attend, care about, and give consideration to oneself and others must fit within relations that involve the relevant forms of mutual deference and so avoid paternalistic and/or domineering modes of

[83] To be clear, I am not assuming there is only one notion of moral danger we might use to evaluate vices. I simply want to provide a concise way of unifying the cluster of moral concerns in light of which we can see vanity as a serious vice, and I think my Rousseuaian account fits the bill. There may be other accounts we could use, but I suspect they would need to similarly draw on some commitment to relational egalitarianism.

[84] This reflects a strong current of recent thought about Rousseau's account of amour-propre – see, e.g., (Dent 1988 and 2006), (Neuhouser 2008), (Cohen 2010), and (Kolodny 2010).

[85] This is because, following (Kurth 2016, pp. 2-3), anxiety involves general (not targeted) risk-minimizing behavior.

[86] This reflects both the genealogical story of social inequality that Rousseau tells in the *Discourse on Inequality* (1997a, pp. 164-72) and his account of the psychological development of young children in *Emile* (1979, pp. 65-7).

[87] See also Neuhouser (2008; 2014) and (Cohen 2010).

interaction. Without such attitudinal traits, persons cannot participate in the egalitarian social practice that solves the problem of inflamed amour-propre.[88]

How, then, should we classify the moral dangerousness of vices on a Rousseauian view? There are two criteria, and they both derive from the ideal of egalitarian, mutually deferential social relations that mitigate inflamed amour-propre. First, a vice is dangerous to the extent that it is immune to standard forms of moral correction (i.e., reproach, critical questions, requests to empathize) by other persons. The sort of mutual deference discussed above involves taking seriously other persons' judgments about how to structure one's interactions with them, including their moral judgments about how they deserve to be treated. The expression of such judgments will often involve some form of moral correction. Insofar as a vicious trait makes one discount or ignore such judgments, it makes one unfit to fully participate in the egalitarian social practice, and it does so in a way that is hard to cure, given that one is unmoved by the standard ways that persons help improve one another. It is thus morally dangerous as an *ineradicable* vice. This is the kind of danger that Rousseau emphasizes in the opening quote, when he describes vanity as the vice that "is the most difficult to destroy."

A related but broader criterion for moral danger is the extent to which a vice is more generally incompatible with participating in egalitarian social relations that mitigate inflamed amour-propre. This is the moral danger of being an *inegalitarian* vice. Vices that are ineradicable will be inegalitarian, but the reverse is not true. A vice might be susceptible to standard forms of correction (and so not dangerous according to the first criterion) while still making its possessor unfit in some other way to participate in an egalitarian practice of mutual deference. For instance, as I discuss below, the form of vanity that I call "entitled conceit" is incompatible with maintaining social relations of equality but still is susceptible to interpersonal moral correction. It is thus morally dangerous as an inegalitarian vice, even though it is eradicable.

You might wonder whether this second criterion of moral dangerousness simply lays out a broad, Rousseauian conception of what a moral vice is. I do not think so, because there may be vices, such as misanthropy, that are compatible with being a participant in egalitarian social relations (and so also are susceptible to interpersonal correction) but that, nevertheless, undermine an individual's moral development as a participant in egalitarian social relations. While bad, such vices will not pose the kind of interpersonal danger captured by the two criteria above.

In sum, on my Rousseauian view, vices are morally dangerous to the extent they are ineradicable (i.e., resistant to standard forms of interpersonal moral correction) and/or inegalitarian (i.e., incompatible with egalitarian social relations). These criteria derive from the idea that morally desirable character traits

[88] This may strike you as a pragmatic or consequentialist account of good character traits, on which good character traits are determined by their desirable social effects. That is true but somewhat misleading, in two ways. First, my account is specific about what kind of desirable social effect good character traits serve – namely, they serve the aim of mitigating inflamed amour-propre. This is not a consequentialist account that focuses broadly on bringing about the best outcomes overall. Second, the desirable character traits that make us fit to participate in an egalitarian practice of mutual deference help constitute, rather than merely cause or make more likely, our solution to the problem of inflamed amour-propre. These traits are a part of what it is to avoid the social conditions that inflame amour-propre. This means that, just as many Aristotelians claim that virtuous character traits help constitute the collective realization of *eudaimonia*, I claim that virtuous character traits help constitute the collective mitigation of inflamed amour-propre. You might further ask: I see why it is important realize *eudaimonia*, but why is it important to mitigate inflamed amour-propre? I think there are several potential answers to this question. One compelling option is to take seriously Rousseau's suggestion in *The Social Contract* that our political society should make us "as free as before," (1997b, p. 50) where the "as before" refers to the state of nature. Failing to solve the problem of inflamed amour-propre amounts to a failure to recover a social form of the naïve freedom Rousseau imagines we would enjoy in the state of nature, and the reverse might be true, too: failing to recover a social form of our natural freedom amounts to a failure to mitigate inflamed amour-propre.

must enable participation in an egalitarian social practice of mutual deference, which is what we need to solve the problem of inflamed amour-propre.

3. With a clearer sense of the moral dangers at hand, let us now consider the different forms of the vanity-vice, so that we can begin evaluating how dangerous they each are. While there may be multiple ways of carving up the vanity-vice, I will focus on two, cross-cutting distinctions that I think best reveal the unique kinds of moral danger that four basic forms of the vanity-vice present. I will contrast my four-fold account of the vanity-vice with the more coarse-grained analyses offered by Bell (2013) and James (2012).

I initially characterized the vanity-vice as unjustifiably expecting a higher form of regard from others than one deserves. Let me say a bit more about what I take "unjustifiably expecting" to mean. First, the expectation of high regard that is central to vanity is normative, in the sense that it is not merely a descriptive believe about receiving high regard but further (or instead) involves the acceptance of norms of interpersonal regard. These are norms of the form that such-and-such attitude of high regard (respect, admiration, deference, etc.) should be directed (and perhaps expressed) toward individuals with such-and-such properties. This means, second, that there are two ways that the vain person's expectation of undeservedly high regard might be unjustified: first, they accept mistaken regard-norms and have no good reason for doing so, and/or they misapply correct regard-norms to themselves and have no good reason for doing so.[89] [90]

Given this broad characterization of the vanity-vice, what are the different forms it can take? The first distinction concerns whether those who are vain care about the correctness of their expectation of high regard. What I have in mind is parallel to the distinction Harry Frankfurt (2005, pp. 54-61) draws between lying and bullshitting. The liar says what is false out of a concern, in part, with what is true. The bullshitter has no such concern. Similarly, one kind of vain person has an unjustified expectation of high regard that rests, in part, on a concern with the correctness of their expectation of high regard, whereas another kind of vain person has no such concern. Let us call the former "conceited" and the latter "smug." It may be tempting to draw this distinction by saying that the conceitedly vain individual is sincere in their vanity, while the smug person is not. But this is incorrect. Both are sincere insofar as they both accept norms that they take to cast themselves as worthy objects of high regard. What separates them is whether they care about the correctness of their regard-expectations, not whether their expectations are genuine.

This does raise a puzzle about the smug person. How could someone have a genuine normative expectation – which largely consists in accepting regard-norms – and be unconcerned about the correctness of their expectation? To explain how, let's apply a general model of norm-acceptance developed by Chandra

[89] To illustrate, think about the two different ways that someone might be vain about their intellectual abilities. They might accept for no good reason bad norms on which persons deserve basic respect to the extent they are smart and, accordingly, take themselves to deserve much more basic respect than most people, who are not as smart as they are. Or, they might accept correct regard-norms on which smart people are due a specific, narrow kind of admiration for their intellect, but since they are not actually very smart and should know this, they wrongly and for no good reason take those norms to cast them as an apt object of such admiration. Note also an important implication of my caveat "for no good reason": While it is necessary for the vanity-vice that someone expects high regard they do not deserve, this is not sufficient. There may be reasonable, non-vicious ways of expecting undeservedly high regard. This raises the important question of when individuals are to blame for accepting incorrect regard-norms. For an account of whether and how the social prevalence of norms can exculpate those who accept them, see (Calhoun 1989).

[90] By "expectation," then, I refer to the application (and resulting mental states) of norms to social contexts.

Sripada and Stephen Stich (2007). They focus on the intuitive idea that "intrinsic motivation" is central to "internalizing" norms:

> "[We] refer to the type of motivation associated with norms as intrinsic motivation. Our claim is that people are disposed to comply with norms even when there is little prospect for instrumental gain, future reciprocation, or enhanced reputation, and when the chance of being detected for failing to comply with the norm is very small." (p. 284)

We are not simply intrinsically motivated to abide by norms; we also are intrinsically motivated to sanction norm-violations. Summarizing the empirical literature, Sripada and Stich note that,

> "… in various experimental situations and experimental games, people will punish others – *at substantial costs to themselves* – for violations of normative rules or a normative conception of fairness." (p. 288)

Could the smug person be intrinsically motivated to follow and enforce the pertinent regard-norms without caring about the correctness of their normative expectation? You might think it depends on how we characterize this intrinsic motivation. Sripada and Stich generically refer to how we intrinsically value abiding by the norms we internalize, whereas Darwall (2006, p. 158) appeals to a distinctively "deontological" form of motivation that characterizes norm-acceptance and is distinct from pursuing what we intrinsically value. But I think we can set that debate to one side. Regardless of whether we understand a vain individual's motivation as just one instance of the broader phenomenon of pursuing what they intrinsically value, or rather as a distinctively deontological form of motivation, it seems that, in principle, they could be intrinsically motivated to follow and enforce norms without being concerned about the correctness of their corresponding normative expectations.[91]

Why might smug individuals be so motivated? The reasons likely are diverse. Perhaps some deep-seeded insecurity attaches them to bad regard-norms that cast them as more estimable than they worry they are. Or perhaps they were raised with a sense of entitlement on which their entire sense of self is founded, so they cannot imagine questioning the regard-norms they accept.

I do want to rule out a couple of reasons why we might accept norms without caring about the correctness of our normative expectations that do not apply to smug persons. Peter Railton (2006) points out that we can accept a norm without endorsing it. This does not seem an apt way to describe the smug, who are characterized, in part, by their wholehearted commitment to the relevant regard-norms. Alternatively, Cristina Bicchieri (2006, ch. 1) claims that we are motivated to abide by the social norms we accept because we expect others to follow them and they expect us to do the same. It seems that persons could be intrinsically motivated to follow (and enforce) regard-norms for this reason without being concerned about the correctness of their normative expectations. But this also does not aptly describe the

[91] Darwall (2006) might reply that, since regard-norms are second-personal, and accepting second-personal norms entails accepting them as legitimate and valid, which in turn entails accepting that they are supported by second-personal reasons, it follows that accepting regard-norms involves being concerned with the correctness of one's normative expectation of regard. But this does not follow. Taking there to be good reasons (second-personal or otherwise) that justify a norm that one accepts does not entail being concerned about whether there are such reasons, or whether the relevant considerations actually support one's normative expectation in the manner one supposes.

smug, since (as I discuss below) they are characterized by a sort of self-satisfaction that insulates them from others' expectations.

It is more helpful to draw a parallel with the idea of "epistemic insouciance" in virtue epistemology. Quassim Cassam (2018) describes epistemic insouciance as a vice that consists, not so much in ignorance of the evidence or facts that are pertinent to one's beliefs, but a "causal disregard" and oftentimes outright "contempt" for such evidence and facts (pp. 2-6). Individuals who are epistemically insouciant do not care about how the relevant considerations bear on their beliefs and thus are unconcerned about the correctness of their beliefs, even though they maintain those beliefs all the same. Similarly, those who are smug genuinely expect to receive high regard (largely by accepting pertinent regard-norms), even though they are unconcerned about whether their normative expectations are correct. They do not care about whether or how the relevant, normative considerations bear on the correctness of their normative expectations. (We should note here that there seem to be different kinds and degrees of failing to care about the relevant considerations.[92])

It may also be useful to contrast Kant's distinction in *KPV* between self-love and self-conceit with my distinction between conceit and smugness. Kant famously notes that respect for the moral law "only restricts … self-love" but "strikes down self-conceit." (5:73) One thing that is immediately apparent here is the slipperiness of the term "conceit." I think my technical usage of "conceit," as distinct from "smugness," reflects some patterns of ordinary usage, but it is difficult to make any philosophical headway without some artificial stipulation. More to the point, I do not think that Kantian self-love entails vanity, which means that Kant's notion of conceit covers both conceit and smugness as I conceive of them.[93] Kant's distinction between self-love and self-conceit is thus closer to a distinction between the vanity-vice and something else, rather than a distinction between different forms of that vice. In *MS*, Kant more closely focuses on the vanity-vice, what he calls "arrogance," including a strain of arrogance that has some features of smugness.[94] But he still does not clearly draw the conceit-smugness distinction I am drawing here and, unlike in the *KPV* passage, he does not highlight what is distinctively morally worrisome about this vice. It is merely one among several vices that are opposed to the proper regard we have for one another.

In addition to the distinction between those who do (conceited) and those who do not (smug) care about the correctness of their expectation of high regard, there is a distinction between the kinds of high regard that vain persons expect. Following Darwall (1977; 2006, pp. 122-6), we can distinguish the vain person who expects a kind of high "appraisal," which is earned or merited along some putative measure of excellence, from the one who expects "recognition" of (and thus deference toward) some presumed

[92] To explain, sometimes the smug rest their normative expectations on a thin rationale that functions, not to satisfy a concern for the correctness of those expectations, but to preempt any serious questions about them. (We will see an example of this below in n. 22, in my discussion of the "tech bro.") In other cases, no such rationale is present. The smug also vary in the extent to which they do not care. For many smug individuals, there may be a threshold of seriousness in the concerns raised about the correctness of their normative expectations, above which they do care and below which they do not, and this threshold may be higher or lower for different smug individuals. It may also be that, for many smug individuals, they only care about certain kinds or sources of reasons that bear on the correctness of their normative expectations, e.g., they care about prudential considerations but not moral ones. While these are important complications, they do not bear directly on the main conclusion I reach below about the uniquely dangerous character of what I call "entitled smugness," and so I will set these complications aside going forward.

[93] This further implies that Kantian respect for the moral law will "strike down" both conceit and smugness, although perhaps in different ways, which seems right.

[94] Kant claims that an arrogant person may be a "*conceited ass*, that is, that he shows an offensive lack of understanding in using such means as must bring about, on the part of others, the exact opposite of his end[.]" (6:465) What Kant calls "an offensive lack of understanding" points toward the normative carelessness that I take to characterize smugness.

authority that is grounded in his capacities, traits, and/or roles. Call the former "arrogant" and the latter "entitled" (in the pejorative sense).[95]

This second distinction is important because, as I explain below, an exaggerated expectation of appraisal fits more easily into egalitarian social relations (of the sort that Rousseau cares about, at least) than an exaggerated expectation of recognition. Those who demand unjustified authority-recognition are demanding deference from others in a manner that is incompatible with maintaining egalitarian social relations.

When we combine the two distinctions, we find four basic forms of the vanity-vice: **(i) arrogant conceit:** someone unjustifiably expects high appraisal they do not deserve, and they care about the correctness of their expectation (of regard), **(ii) entitled conceit:** someone unjustifiably expects deferential recognition they do not deserve, and they care about the correctness of their expectation, **(iii) arrogant smugness:** someone unjustifiably expects high appraisal they do not deserve, and they do not care about the correctness of their expectation, and **(iv) entitled smugness:** someone unjustifiably expects deferential recognition they do not deserve, and they do not care about the correctness of their expectation.

It is illuminating to see how this four-fold account of vanity overlaps with, but still departs from, Bell's (2013) account of "superbia" and James' (2012) account of "assholes." Bell describes superbia as the vice of believing that one has comparatively high status, desiring that this status be recognized, and, in so believing and desiring, manifesting ill will. (Bell 2013, p. 109) Our accounts are not so far apart at a first pass. Where I place weight on the idea of unjustifiable normative expectations, Bell depends on the idea of ill will. If we apply the Darwallian distinction between appraisal and recognition (which Bell uses elsewhere (e.g, ibid., pp. 170-1)) to distinguish between appraisal-meriting and recognition-meriting forms of high status, we arrive at something like my distinction between arrogance and entitlement. And if we apply Cassam's (2018) notion of epistemic insouciance to the superbic agent's belief in her high status, then we arrive at something like my notion of smugness. However, Bell's account of how we should respond to superbia does not make room for this notion of smugness. She thinks that the attitude of contempt is the appropriate response to superbia, and it curbs superbia by presenting "its target as having a comparatively low status … in virtue of their superbia." (Bell 2013, p. 128) That is, contempt draws the superbic agent's attention to the inferior moral status he earns through his superbia. If we take seriously my distinction between conceit and smugness, it is clear that contempt will not curb smug vanity. As I discuss below, it is only if someone cares about the correctness of their inflated normative expectation that they are susceptible to change in light of being shown how their expectation is incorrect. The smug have no such concern.

For James (2012, p. 12), an asshole is someone who "systematically allows himself to enjoy special advantages in interpersonal relations out of an entrenched sense of entitlement that immunizes him against the complaints of other people." This is both wider and narrower than my conception of the vanity-vice. It is wider, because "special advantages" may include other things than the forms of high regard that the vain expect to receive. It is narrower, because not all forms of the vanity-vice involve being "entrenched" against others' complaints. If we focus only on what I call smugness, and if we narrow James' account to focus on high regard (and not the wider class of special advantages), there still remains an important difference between my account of smugness and James' account of being an asshole. Someone might be immune to others' complaints about their inflated expectations of high regard and still care about the correctness of

[95] To illustrate, someone who thinks they are a much better judge of moral character than they really are and expects to be *admired* as such is vain about appraisal, and so, on my usage, is arrogant. By contrast, if they take their capacity for character-assessment to ground some authoritative standing to make pronouncements that other people must *heed*, then they are vain about recognition and so, on my usage, are entitled.

those expectations, and so they might be an asshole without being smug.[96] This matters, because the asshole who cares about the correctness of their inflated expectations of high regard is still open, in principle, to revising those expectations in light of being shown how the expectations are incorrect (of course, this demonstration will have to come through some other means than others' moral complaints). The same is not true for the smug person. Their immunity to complaints is just one symptom of their broader lack of concern about the correctness of their regard-expectations.

4. We are now in a better position to begin evaluating the dangerousness of the different forms of the vanity-vice. Using my Rousseuaian framework, there are two main questions we need to ask about each of its forms: Is it eradicable by standard forms of interpersonal moral correction,[97] and is it compatible with participating in (the relevant kind of) egalitarian social relations?

Let us begin with arrogant conceit, which I think is the most benign form of vanity. (This is faint praise, to be sure.) On its own, arrogant conceit is eradicable by interpersonal moral correction. This is because the arrogantly conceited care about the correctness of their expectation of high regard (they are conceited, not smug), and the kind of high regard they care about receiving does not entail demanding deference from others, at least not in any way that would preclude taking seriously the moral correction offered by others. Of course, whether they actually do so will depend on myriad factors – how the arrogantly conceited person is related to the individuals offering correction, how the correction is expressed, and the source of their arrogant conceit. But all else equal, there is nothing about arrogant conceit that resists revision in the face of interpersonal moral correction.[98]

Arrogant conceit is also generally compatible with participating in egalitarian social relations. Unjustifiably expecting admiration along some measure of excellence is not, by itself, in tension with maintaining relations of mutual deference with those one takes to have the same basic standing as oneself to govern their lives and interactions. That is, unjustifiably expecting admiration as a great pianist, or astute judge of character, or math whiz, does not entail rejecting other people's standing to govern their lives or weigh in about the structure of one's interactions with them, or even thinking that one has higher standing to govern one's life and interactions than they do. This is not to say that people do not often slide from arrogant conceit into some other form of vanity that does disrupt their ability to relate to others as equals. But there is nothing about arrogant conceit itself that is incompatible with egalitarian social relations (at least those that mitigate inflamed amour-propre).

Like arrogant conceit, entitled conceit is open to moral correction, with an important qualification: if the kind of authority-recognition that one expects (in virtue of one's entitled conceit) concerns the authority to settle moral questions, then one's vanity will make one resistant to receive moral correction, at least from those who do not recognize and/or do not share one's wrongly presumed high moral authority.

---

[96] I should note that I have no qualms with James' appropriation of the terms "smug" and "smugness" (ibid., pp. 39-43), but it is clear that he does not use these terms to capture the important distinction I want to draw here between conceit and smugness.

[97] Above, I claimed that smugness is importantly different from being an asshole in James' sense because the normative carelessness of smugness extends beyond its immunity to the moral complaints of others. Am I here walking back this claim, in focusing on the danger of being immune to interpersonal moral correction? No, because interpersonal moral correction extends beyond moral complaints, including such things as critical questions and requests to empathize. James (ibid., pp. 25-6) purposefully focuses only on the kinds of interpersonal confrontation that fit into the cast of second-personal accountability. There are many important forms of correction that figure into our interactions as equal persons that are not expressions of second-personal accountability.

[98] It is worth noting that not all forms of arrogant conceit may be morally vicious and, accordingly, not all forms of correction directed at it may count as moral correction.

In this case, even though one cares about the correctness of one's expectation of high regard and so, in principle, is open to revising one's expectation in light of reasons to reject it, one will discount such reasons when offered by those who fail to recognize one's presumed standing and/or are not taken to share it. Otherwise, entitled conceit is susceptible to interpersonal moral correction for the same basic reason that arrogant conceit is, stemming from the fact that the conceited care about the correctness of their inflated expectations of high regard.[99]

Entitled conceit is worse than arrogant conceit along the second dimension of moral danger, being inegalitarian. This is because any form of entitlement (as described above) unjustifiably expects undue deference. If one unjustifiably expects undue deference, then one is expecting authority-recognition in a manner that is incompatible with upholding social relations of basic equality.[100] Why think that? I am not sure there is theory-neutral answer to this question. From the broadly Rousseauian perspective I take here, the basic criterion of a justifiable expectation of deserved authority-recognition just is the compatibility of this expectation with social relations of equality (as specified above). That is, are these expectations based on norms that support the mutual deference of social relations of equality, and do the agent's reasons for holding these expectations cohere with her participation in social relations of equality? (These two questions address the two ways that an agent's normative expectation of undeservedly high regard might be unjustified, as described above in section 3.) When an agent exhibits entitled conceit, the answer to one or both questions will be "no," because the agent will be applying norms that call for authority-recognition that is incompatible with the mutual deference of social relations of equality, and in doing so, she likely will be responding to considerations that prevent her full participation in egalitarian social relations.

Let us now turn to the two forms of smugness. Arrogant smugness will be more morally dangerous than either form of conceit insofar as it is generally immune to moral correction. If someone unjustifiably expects appraisal-regard for some form of perceived excellence but does not care about the correctness of their expectation, then they will be unmoved by the concerns expressed by others about this expectation (barring some further motive for registering these concerns). They might be bothered and even offended if the expression of these concerns prevents them from receiving the appraisal they expect, but this will not

[99] You might wonder whether conceit (either arrogant or entitled) is as correctable as I have suggested. Following James (2012, pp. 27-8), you might point out that the conceited are often "entrenched" in their expectation of high regard and, accordingly, are not open to adjusting their expectations in light of other persons' moral concerns. But I think much hangs on why their inflated expectations are entrenched. When we examine the reasons for their entrenchment, we often see that the entrenchment does not follow from their conceit. Sometimes the conceited have entrenched expectations because central aspects of their self-conception depends on sustaining their expectation of high regard. If so, then it is not their conceit, but how their self-conception depends on their conceit, that makes them resistant to interpersonal moral correction. Other times they have entrenched expectations because, as I discussed above, they are morally entitled, i.e., they expect deference to their standing to issue moral judgments. This is a narrow form of conceit and so does not pose any problems for my claim that most forms of conceit are susceptible to interpersonal moral correction. And in yet other cases, being conceited is conjoined with having a dismissive attitude toward others by which one discounts or ignores their moral concerns. This is the kind of dismissive entrenchment on which James (ibid.) focuses. But such dismissiveness does not follow from being conceited (whether in an arrogant or entitled manner) and so does not entail that my claim – about the basic correctability of conceit – is incorrect.

[100] This does not imply that authority-recognition is justifiably expected only if it concerns the recognition of equal (i.e., non-hierarchical) authority. For instance, it is plausible that parents justifiably expect their adolescent children to recognize their unequal parental authority, or that the proper operation of many organizations requires that decisions are made by managers whose hierarchical organizational authority is recognized by their subordinates (keeping in mind the organizational constraints proposed in (Anderson 2017, pp. 66-71)).

lead them to doubt or scale back their expectation.[101] To return to James' idea of entrenchment, this is a kind of entrenchment that is essential to the vice itself.

When we look at the second criterion of moral danger, however, arrogant smugness is less dangerous than entitled conceit. This is because, as with arrogant conceit, it is not generally incompatible with egalitarian social relations, at least not beyond its resistance to registering others' concerns, whereas the authority-recognition sought by entitled conceit is deeply inegalitarian.

Entitled smugness is the worst, compared not just to arrogant smugness, but to all the other forms of the vanity-vice. It is immune to interpersonal moral correction, given that it involves not caring about the correctness of one's expectation of high regard (barring, as with arrogant smugness, some separate motive to worry about others' concerns about one's regard-expectation). And it is more broadly incompatible with social relations of equality, given that it involves an unjustifiable expectation of undeserved authority-recognition. This is the form of vanity that truly fits Rousseau's description of being "most to be feared" and "most difficult to destroy."[102]

To illustrate and conclude, it might be helpful to consider a familiar kind of entitled smug character. Consider the "tech bro" of Silicon Valley, a type made infamous by Justin Keller, who describes himself as "entrepreneur, music lover, beer connoisseur, sports enthusiast, traveling the world." (https://justink.svbtle.com) In February 2016, Keller penned an open letter to then-San Francisco Mayor Ed Lee and Police Chief Greg Suhr, in which he complained about various (relatively harmless) encounters he had with (putatively) homeless individuals in San Francisco. He summarized his argument as follows:

> "The wealthy working people have earned their right to live in the city. They went out, got an education, work hard, and earned it. I shouldn't have to worry about being accosted. I shouldn't have to see the pain, struggle, and despair of homeless people to and from my way to work every day. I want my parents when they come visit to have a great experience, and enjoy this special place." (Ibid.)

While there are many obvious criticisms to make here, I want to focus on two aspects of the attitude Keller expresses that I think aptly illustrate my notion of entitled smugness. First, Keller unjustifiably takes himself to occupy a special socio-economic position – one of a "wealthy working" person – that commands undeserved deference from others – in particular, he demands that those who manifest symptoms of homelessness, substance addiction, and/or mental illness not disturb his sensory field. Second, Keller's expression of his normative expectation casually disregards relevant normative considerations, insofar as he is utterly impervious to the countervailing claims that the dispossessed of San Francisco might make on him. He (and his kin) deserve "to have a great experience" (and, more broadly, to have the world cater to their enterprising desires[103]), and he seems to have no concern about whether or how he should question this overblown normative expectation. In short, he unjustifiably expects the deference he thinks is due a

[101] How can the smug be reformed, then? I am not sure, but I would guess that it requires addressing the psychological underpinnings of their inflated normative expectations, which, as I mentioned above, may include such things as deep insecurities that are not consciously or rationally related to the normative expectations themselves.

[102] You might think this is trivially true, because entitled smugness appears to be a manifestation of inflamed amour-propre. But actually, while entitled smugness disrupts anti-inflammatory social relations, it is not plausibly an expression of inflamed amour-propre, at least not directly. Inflamed amour-propre is borne out of an anxiety about being taken seriously by others. Those who are smug tend not to be anxious; they are "above" being anxious about what others think about them, even though they normatively expect others to pay them heed.

[103] See Chang's (2018) description of Silicon Valley's "brotopia": "… [F]ounders think they can change the world. And they believe that their entitlement to disrupt doesn't stop at technology; it extends to society as well."

wealthy working person, and he does not care about the correctness of this expectation.[104] He manifests entitled smugness. (I say "manifests" to emphasize that I am drawing inferences only on the basis of what Keller writes in his open letter, and not on any other evidence about his character. I certainly do not know the bro.)

Some of what is maddening about the attitude Keller conveys reflects the two dimensions of the unique moral dangerousness of entitled smugness. First, we feel that he is unjustifiably elevating himself above us (provided that we do not satisfy the nebulous criteria for being a wealthy working person) and that, in doing so, he takes himself to have a kind of authoritative standing that we must heed. Second, we sense that there is no way we could lodge our legitimate concern in his mind. Even if we were a fellow wealthy working person (which probably entails being a man[105]), he would likely brush us off with a characteristic "Whatever, bro."[106] ("Do you even program?") While all forms of vanity may deserve some rebuke, entitled smugness is unique in the moral danger it contains.

REFERENCES

Anderson, E. (2017). *Private Government: How Employers Rule Our Lives (and Why We Don't Talk about It).* Princeton: Princeton University Press.

Bell, M. (2013). *Hard Feelings: The Moral Psychology of Contempt*. New York: Oxford University Press.

Bicchieri, C. (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms.* Cambridge: Cambridge University Press.

Calhoun, C. (1989). Responsibility and Reproach. *Ethics*, *99*(2), 389-406.

Cassam, Q. (2018). Epistemic Insouciance. *Journal of Philosophical Research,* (in press).

Chang, E. (2018, February). "Oh My God, This is So F---ed Up": Inside Silicon Valley's Secretive, Orgiastic Dark Side. *Vanity Fair*. Retrieved from: https://www.vanityfair.com/news/2018/01/brotopia-silicon-valley-secretive-orgiastic-inner-sanctum.

Cohen, J. (2010). *Rousseau: A Free Community of Equals*. Oxford: Oxford University Press.

Dent, N. (1988). *Rousseau: An Introduction to his Psychological, Social, and Political Theory*. Oxford: Basil Blackwell.

_____ (2006). *Rousseau*. New York: Routledge Press.

Darwall, S. (1977). Two Kinds of Respect. *Ethics*, *88*(1), 36-49.

[104] He does acknowledge that "people are frustrated about gentrification," and he apologizes for using the term "riff raff" in his letter. But he does not consider what the frustrations about gentrification actually are, nor does he show any understanding of why or how using "riff raff" is objectionable. He merely concedes that it is "insensitive," as though the individuals designated by "riff raff" and their allies are too easily offended.

[105] See ibid, describing the sexist practices that prevail in Silicon Valley: "Great companies are built in the office, with hard work put in by a team. The problem is that weekend views of women as sex pawns and founder hounders can't help but affect weekday views of women as colleagues, entrepreneurs, and peers."

[106] But might there be a form of entitled smugness *qua* tech bro that would allow for registering a concern leveled by a fellow bro? This raises the important question of how localized smugness might be. Could an individual be smug only within certain contexts, institutions, or groups of individuals? I see no reason why not, especially in light of the variations in how smug individuals might fail to care about the correctness of their regard-expectations – see n. 10 above.

_____ (2006). *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, MA: Harvard University Press.

Frankfurt, H. G. (2005). *On Bullshit*. Princeton: Princeton University Press.

James, A. (2014). *Assholes: A Theory*. New York: Anchor Books.Kennedy, Kim & Strudler (2016)

Kant, I. (1996a). *The Critique of Practical Reason*, in *The Cambridge Edition of the Works of Immanuel Kant: Practical Philosophy*, trans. and ed. M. Gregor. Cambridge: Cambridge University Press.

_____ (1996b). *The Metaphysics of Morals*, in *The Cambridge Edition of the Works of Immanuel Kant: Practical Philosophy,* trans. and ed. M. Gregor. Cambridge: Cambridge University Press.

Keller, J. (2016, February 15). Open letter to SF Mayor Ed Lee and Greg Suhr (police chief). [Blog post] Retrieved from: https://justink.svbtle.com/open-letter-to-mayor-ed-lee-and-greg-suhr-police-chief.

Kolodny, N. (2010). The Explanation of Amour-Propre. *The Philosophical Review*, *119*(2), 165-200.

Kurth, C. (2016). Anxiety, Normative Uncertainty, and Social Regulation. *Biology and Philosophy*, *31*(1), 1-21.

Neuhouser, F. (2008). *Rousseau's Theodicy of Self-Love: Evil, Rationality, and the Drive for Recognition*. Oxford: Oxford University Press.

_____ (2014). *Rousseau's Critique of Inequality: Reconstructing the Second Discourse*. Cambridge: Cambridge University Press.

Railton, P. (2006). Normative Guidance. *Oxford Studies in Metaethics*.

Rousseau, J. J. (1979). *Emile, or On Education*, trans. A. Bloom. Basic Books.

_____ (1997a) *Discourse on the Origin and the Foundations of Inequality among Men.* In *The Discourses and other early political writings*, ed. V. Gourevitch. Cambridge: Cambridge University Press.

_____ (1997b). *The Social Contract*. In *The Social Contract and other later political writings*, ed. V. Gourevitch. Cambridge: Cambridge University Press.

Rozeboom, G. J. (2018). The Anti-Inflammatory Basis of Equality. *Oxford Studies in Normative Ethics*.

Sripada, C. & Stich, S. (2007). A Framework for the Psychology of Norms. In *The Innate Mind: Culture and Cognition*, vol. 2 (pp. 1-34). Oxford: Oxford University Press.

Addiction as Evidence:
Frankfurt's Unwilling Addict and the Explanatory Gaps of Mesh Theories of Responsibility

Cami Koepke

Mark initially started using oxycodone by medical prescription as he recovered from an injury resulting from an IED explosion in Iraq.[107] The pain from the injury, though, didn't subside and he eventually switched to heroin as a more cost effective pain killer. After recognizing he had become addicted to opioids and not wanting to be an addict, he devised a plan to detox by using small doses of heroin and Suboxone, a drug that lessens the efficacy of opioids by dampening cravings and withdrawal symptoms. While implementing his plan, though, a dealer sold him heroin laced with the considerably more powerful fentanyl. Trapped by a level of pain and wanting he had never before experienced, with Suboxone doing nothing to ease the wracking aches and fierce cravings, Mark felt at the mercy of an addiction he didn't want and couldn't overcome. Despite desiring drugs, he hates that he is motivated to keep using. What he wants and what he wants to want are at war.

Mark might be described as an "unwilling" addict because he does not identify with his motivating desires for drug use. Addicts like Mark have influenced moral responsibility theorizing since Harry Frankfurt's (1971) hypothetical cases of the unwilling and willing addicts. In contrast to the unwilling addict, the willing addict identifies with his motivating desires for drug use. The cases serve a two-fold role in theorizing. Intuitions about the two addicts are thought to serve as evidence for the conditions at the heart of responsibility assessments. But also a theory's explanatory robustness is in part measured by whether it can offer a principled explanation for intuited reactions.[108]

Historically, the cases of the willing and unwilling addicts have been thought to support "mesh" quality of will views. For quality of will views in general, the foundational feature of concern in ascribing blameworthiness is an agent's evaluative orientation toward the well-being or rights of others. Mesh theories look specifically at the structural fit or "mesh" between an agent's evaluative orientation and her motivating attitudes that effectively drive her action.[109] Agents are responsible for actions that result from the right structural fit, i.e. an agent's evaluative orientation coheres with her motivating attitudes that drive her action. These theories consider this fit to be both a necessary and sufficient condition for wrongful actions to be assessed as blameworthy. In a nutshell, the willing addict's evaluative orientation meshes with his addictive motivation so he is more blameworthy for what he does than the unwilling addict whose evaluative orientation does not.

What the case of the willing addict illustrates and supports about the conditions of responsibility has received extended attention.[110] Q*uality of will* theorists contend the case shows that identification is the key condition for responsibility, while *reasons-responsiveness* theorists contend that control over normative capacities plays this role. In contrast, little is said about the unwilling addict though both sides freely use the case in their defense.

Here I will consider the unwilling addict and argue that when we attend more closely to the role that effort plays in how the case is described, it becomes apparent that appealing to identification will not provide the means to explain important distinctions between relevantly different types of addicts and akratic actors. To defend this claim, I will first look at Frankfurt's original iteration of the willing and unwilling addicts, showing how these cases are typically thought to support mesh theories of

[107] Percy, Jennifer. (10 Oct 2018) Trapped by the 'Walmart of Heroin'" in The New York Times Magazine. *https://www.nytimes.com/2018/10/10/magazine/kensington-heroin-opioid-philadelphia.html*. Accessed 10 Oct. 2018.
[108] For some such explanations, see Fischer & Ravizza (1998, pp. 35, 69, 74); Wolf (1980, p. 155); Arpaly and Schroeder (2014, pp. 274-289).
[109] I follow Fischer and Ravizza (1998, p. 184) in using this name for these view. This sort of view is also referred to as a "structural" or "hierarchical" theory of responsibility (Talbert, 2016 p. 85) and are classified by some among the broader category of "identificationist" views (Jaworska, 2016).
[110] e.g. Leon (2001) and Sripada (2017).

responsibility. In the second section, I will show how the unwilling addict's concerted effort to fight his errant motivation plays a significant role in intuitive responses to the case. But, further, while effort can signal a high degree of identification, weak or absent effort provides no guidance for identification's strength. This explanatory gap leads mesh views to assess many unwilling addicts the same as akratic actors. In the third section, I will show that attending to two ways in which addicts might be thought to possess or lose control over their addictive desires provides the means to distinguish between relevantly different types of addicts and akratics. Making these distinctions suggests that theories of blameworthiness that appeal to control over abilities have more explanatory robustness than theories that attend to agential identification.

I. Mesh Theories, Blameworthiness, and Addiction

a. The Basics

Mesh views have had an active life in discussions on agency and responsibility since Frankfurt's 1971 seminal essay "Freedom of the Will and the Concept of a Person." In this work, cases of addiction are thought to show that an agent can act freely act despite acting on compulsive motivating desires because acting freely depends solely on the details of an agent's internal psychological structure. Addiction's assumed compelling nature makes it illustratively apt to show relevant psychological variations in the relationship between higher-order attitudes – what an individual wants to want – and his effective motivating desires – what the individual is actually motivated to do. An addict can identify with or withhold his identification from his effective motivation to use drugs. The two addicts might be characterized like this:

> Willing Addict: This addict has conflicting motivating desires for and against using drugs, but identifies only with his motivating desires to use drugs. Though compelled to use drugs, he identifies with this desire that effectively results in action.
>
> Unwilling Addict: This addict has conflicting motivating desires for and against using drugs, and identifies only with his desire *not* to use drugs. Compelled to use drugs anyway, he acts despite what he wants to want to do.

Frankfurt contends that an addict acts freely if and only if he identifies with his effective motivating desire, regardless of whether he controls the efficacy of the first-order desire that actually motivates his action. Key here is that the identification via a higher-order desire explains – or fails to explain – why the person acts as he does.[111] The willing addict's identification with his effective motivating desire is what distinguishes him from the unwilling addict.

On Frankfurt's mesh theory, the agent's evaluative orientation is seen as a *desiderative* attitude, and the addicts are described with this sort of psychological architecture in mind. Frankfurt's addicts, however, might also be thought to illustrate and support a *cognitively-based* mesh theory like the sort articulated by Gary Watson (1975).[112] Watson argues that the act of identification as performed by *evaluative judgments,* rather than higher-order desires, best represents an agent's moral concern. Watson conceives of evaluative judgments as comprised by beliefs and values held in a "cool and non-self-deceptive moment" that regard the practical ends an individual thinks make for a good and fulfilling life (p.215). Motivating desires are ultimately value neutral regarding their targets. Values and motivation can work together, e.g. we can value what we desire or come to desire what we value. But the two can also come apart, such as when one is effectively motivated to do something she desires but does not value. Free action follows when evaluative judgments and motivation have the right relationship.

---

[111] When identifying what explains the willing addict's action, we would say that he acts because of his identification with the effective desire to take the drug as well as because of the compelling nature of this desire. As Frankfurt (1971) clarifies in the discussion of the willing addict: "His will is outside his control, but, by his second-order desire that his desire for the drug should be effective, he has made this will his own. Given that it is therefore *not only because of his addiction that his desire for the drug is effective*, he may be morally responsible be morally responsible for taking the drug" (p.20, emphasis mine).

[112] It should be noted that Watson moves away from this account in his later work.

Rather than making a case for one mesh view over the other, in this paper I will speak more inclusively of an agent's identification with her motivating desires, leaving open as to whether identification is best understood in terms of desires or evaluative judgements.

b. Clarifications

The distinction and possible conflict between the motivation that one identifies with and what one is actually motivated to do serves as a crux for Frankfurt's notions of free action and agency as well as for his theorizing about the conditions of responsibility. According to Frankfurt, assuming that an individual meets the minimum standards for agency by identifying with motivating desires, the agent is deemed morally responsible for action that is done freely. Though Frankfurt (1971) and early Watson (1975) say little more than this in terms of moral responsibility, we might build a mesh theory of blameworthiness from these resources. To clarify, I will be concerned with the conditions for blameworthiness but not for praiseworthiness. My reason for this focus is that when discussing the actions of addicts, we're typically concerned with whether blaming is justified.[113]

It's also important to clarify what addicts might be blameworthy for. While Frankfurt's original cases concern drug use, it's contentious whether *drug consumption* in and of itself is a morally wrongful action.[114] However, I think we can see the act of perpetuating one's own addiction as wrongful when the addiction directly results in a neglect of one's moral obligations.[115] Consider David Carr's experience that he recounts in his autobiography *Night of the Gun.* Addicted to intravenous cocaine and experiencing intense cravings, Carr bundled up his infant twins in their snow suits and left them unattended in the car in sub-freezing weather for hours while he met with his drug dealer. The wrongfulness of Carr's action can be variously construed, such as unnecessarily risking considerable harm or failing to attend to what is of utmost moral importance. But it is clear that prioritizing the service of his addiction over fulfilling his parental responsibilities was wrong.

I'll assume that the addicts in my examples developed their condition non-culpably, e.g. by becoming addicted following medical orders or during childhood when they lacked the capacity for appreciating the risks of addiction. This assumption will set aside the complication that an addict is blameworthy for perpetuating an addiction just because she is culpable for becoming addicted in the first place.

Finally, given recent critiques of how philosophers tend to characterize addiction, it's worth saying something about Frankfurt's depiction of the condition. Traditionally, philosophers have tended to view addiction as a purely volitional condition marked by irresistible desires. As Hanna Pickard (2015) and others have argued or noted, though, understanding addiction as involving *literally irresistible* desires is misguided.[116] With the recent critiques in mind, I will assume in the cases I discuss that the addictive desires are not literally irresistible but that the addicts have an extreme inability to overcome their addictive desires. While I do not commit to the idea that such an extreme inability characterizes all cases of addiction, it's plausible to think that at least some cases meet this standard.

c. Blameworthiness

On a mesh theory of responsibility, an agent is blameworthy for what he does when he identifies with the motivating desire that he acts upon. It's plausible to think that the degree of blameworthiness

[113] My focus is on the conditions for assessing an agent as blameworthy for a wrongful action, and I will remain open about the concept of responsibility itself, such as whether it is essentially retributivist in nature. Furthermore, the notion of blameworthiness I have in mind is the sort typically associated with accountability rather than attributability, following the influential distinction made by Watson (2004). Though the distinction is variously interpreted, here assessing someone as blameworthy in the accountability sense means deeming that the person is an appropriate target for blaming responses involving sanctioning, while being blameworthy in the attributability sense means attributing a moral fault to the agent without further thinking sanctions are justified.

[114] See Douglas Husack (especially 2004 but also 1999 and 2000) for an interesting discussion on the possible moral and legal wrongdoing of drug consumption as opposed to the condition of being addiction or the wrongdoing that might follow from either.

[115] Thanks to David Brink for suggesting this sort of case.

[116] See Mele (2004) for another example.

corresponds to the degree that the agent identifies with his effective motivation. So, the willing addict who does what he wants but also identifies with this motivation is more blameworthy than the unwilling addict who does what he wants, in one sense, but does not identify with this effective motivation.

Building from the mesh theory approach to free action and responsibility, we can state the basic idea of blameworthiness and excuse as follows:

> Blameworthiness: an agent is blameworthy for an action if and only if the action is an instance of wrongdoing and the agent identifies with the motivating desire that drives the action.
>
> Excuse: an agent's blameworthiness for an action is excused in whole or in part if and only if the action is an instance of wrongdoing and the agent does not identify in whole or in part with the motivating desires that drive the action.

It's meant to be intuitive that the unwilling addict is worthy of some degree of excuse.[117] It's more controversial, though, whether Frankfurt's willing addict intuitively acts responsibly, and this verdict has recently received a more sustained defense by philosophers like Chandra Sripada (2017). As people like Mark Leon (2001) argue, a case can be made that the primary explanation for the willing addict's drug use just is his compulsive desires rather this his identification with the desires, thereby undermining Frankfurt's project. Here I will set aside this debate and grant that the willing addict as described in the original case is intuitively more blameworthy than the unwilling addict.

II. Identification and the Role of Effort

In this section I argue that the unwilling addict's concerted *effort* to resist his errant desires does significant work in making him seem worthy of some degree of excuse for his wrongdoing. A mesh theorist might say that the concerted effort indicates a high degree of identification against the addictive desires, but I will argue that this response has explanatory limitations that become quickly apparent when we consider other types of unwilling addicts.

Unwillingness and Identification

The idea that the unwilling addict is worthy of excuse because of his psychological unwillingness has initial appeal. Most people at times have felt a degree of alienation from a desire and its related behavior. It is perhaps this widely shared feeling of psychological alienation that makes Frankfurt's unwilling addict so seemingly worthy of excuse. Note that the unwilling addict does not merely psychologically disapprove of his addictive desires, though. Rather, Frankfurt emphasizes the addict's concerted efforts to resist these desires. It is this effort that makes the deservingness of an excuse all the more intuitively compelling. Consider Frankfurt's description:

> "[the unwilling addict] hates his addiction and always struggles desperately, although to no avail, against its thrust. He tries everything that he thinks might enable him to overcome his desires for the drug. But these desires are too powerful for him to withstand, and invariably, in the end, they conquer him. He is an unwilling addict, helplessly violated by his own desires" (1971, p.12)

The unwilling addict's effortful struggle to fight his addictive motivating desires is key to communicating his unwillingness. When he acts wrongly, the fact that he's exhaustively tried all of his options makes him especially intuitively compelling. To appreciate the role that effort plays in our intuitions, compare the description to an addict whose lack of identification is merely psychological:

> Aspirational Unwilling Addict: This addict is compelled by motivating desires to use drugs, but only identifies with her desires not to use drugs. She always regrets getting high and feels deeply ashamed and vows never to use again when she finally succumbs to the desire. This addict psychologically disapproves of her addictive desires but takes little action to try to quit, foregoing the various options available to her.

[117] See Fischer (2012, p.129-131); Fischer & Ravizza (1998, p. 82); Wolf (1993, p.28); and Scanlon (1998, p. 290), and Jaworska (2017, p.19) for some examples.

Intuitively, the original unwilling addict and the aspirational unwilling addict seem importantly different in terms of blameworthiness. This difference looks to turn on effort, but how might mesh theories explain effort's importance? If the presence or absence of psychological identification is what really matters for assessments, why does effort matter?

One possible response on behalf of mesh theories is that effort reflects the degree of an actor's identification with a motivation either by constituting or by signaling the degree of her unwillingness. The more effort an addict exerts to resist an action, the more psychologically unwilling she is when she performs the action anyway. As this response might go, while both the original and the aspirational unwilling addicts are worthy of *some* excuse on account of their unwillingness, the original addict is far more unwilling as signified or constituted by her effort, and is thus deserving of a greater degree of mitigation.

In fact, a response of this sort might be what mesh theories need to respond to another worry commonly raised about their extension of excuse.[118] This complaint concerns how mesh theories handle akratic actors. We can picture the standardly conceived akratic actor as this:

> Standard Akratic: This person acts on a motivating desire to do an action X, but does not identify with the motivating desire.

As the worry goes, if the unwilling addict is worthy of excuse on account of a lack of identification, then akratic actors who likewise don't identify with their effective motivation are equally deserving of excuse. Mitigating blameworthiness for akratic actors strikes many as a promiscuousness extension of excuse because, unlike in the case of addiction, akrasia does not involve an extreme inability to resist errant motivating desires. With an appeal to the feature of effort, though, a mesh theorist can say that there is a responsibility relevant feature that distinguishes the original unwilling addict from either the aspirational addict or the akratic actor. Namely, the original unwilling addict exerts far more effort, indicating a much greater degree of unwillingness. The aspirational addict's middling effort reveals only a half-hearted withdrawal from his errant desires for drugs, as does the akratic actor's level of effort.

This mesh response looks initially promising, but the explanation becomes less compelling when we look more carefully at the relationship between effort and identification. One option is to say that effort is a *constitutive part* of identification; the greater the degree of effort, the greater the degree of identification for or against some effectively motivating desire. If this is the relationship, then the degree of effort and the degree of identification for or against a motivating desire are positively correlated.

However, I think we should resist thinking that effort constitutes identification because the two can come apart. For instance, consider cases in which a *lack of effort* belies an agent's psychological attitudes. Take Owen Flanagan's (2017) case of the "resigned" addict. As Flanagan describes this addict, she genuinely tried to stop using by accepting all offers of help, etc., but after repeatedly failing has finally given up actively resisting the addictive motivating desires that she was unable to overcome. We might picture her like this:

> Resigned Unwilling Addict: This addict is highly motivated to use drugs, but does not identify with these desires. This addict once tried everything to resist using, but after years of repeated failure has finally given up trying. She still always deeply regrets getting high. Every time she uses, she feels deeply ashamed and is decidedly against her drug related desires and behaviors. Yet, she has become motivationally depleted when the cravings surge and no longer takes action to resist the desires.

The resigned addict has a complicated internal life but represents one plausible reaction to motivating desires that cannot be resisted. She identifies with her desires not to use drugs, but when faced with a nearly insurmountable obstacle of resisting addictive desires her lack of self-efficacy results in a sense of

[118] Talbert (2008, p. 6), Fischer (2012, p.129-131), Jaworska (2017, p.19), and Brink (forthcoming) mention a worry of this sort.

defeat that inhibits her resistance to errant motivating desires.

If my analysis of the resigned unwilling addict is right, then effort is not a constitutive part of identification. It is still open, though, for the mesh theories to say that effort can *signify* an agent's degree of identification even if this signal is imperfect. But such an imperfect signal limits its use for understanding the degree of identification, which means that a mesh theory needs to rely on other features as well to communicate the degree of an unwilling actor's unwillingness. Some possibilities might be feeling shame or regret when one succumbs to motivation with which one does not identify. Again, though, these features are likely imperfect signals. The akratic actor might equally feel great shame or regret after succumbing to a second slice of cake and the truly unwilling addict might become numb to the feeling of failure.

The mesh theorist might be able to fill out the details of psychological identification in such a way that addresses the variety of cases and gives a plausible story about the many ways in which an agent might identify with or withdraw identification from an effective motivating desire. The larger point here, though, is that the feature of effort does much of the work to make the original unwilling addict seem especially deserving of some degree of excuse, but effort is not quite the theoretical resource for mesh theories that it initially appears to be. I think these explanatory limitations motivates a second look at the cases of the unwilling addicts to see what else can do the explanatory work that identification cannot.

III. Addiction, Akrasia, and Abilities

In this section, I argue that the feature of control over volitional abilities provides the theoretical resources to distinguish and explain intuitions to three cases considered above, but only when two ways in which addicts might be said to have or lack volitional control are specified.

Conquering vs. Circumventing

Revisiting the case of the original unwilling addict, another plausible explanation for our intuitive inclination to ascribe him some degree of excuse is due to his inability to overcome his motivating desires. On this interpretation, the original addict's effort to resist his errant desires is significant to our assessments because it shows a reduction in his ability to do what he wants to do. Akin to the relationship between effort and identification, though, effort is only an imperfect signal of an agent's degree of control. While failed concerted effort to resist a motivating desire signals a low degree of control over that motivating desire, the case of the resigned addict shows that absent effort does necessarily signal a high degree of control.

However, unlike the mesh theorist's explanation that I explored above, the proponent of a control view still has considerable resources to draw principled distinctions in cases where concerted effort is missing. But compare again the cases of the original unwilling addict and the aspirational addict. There seems to be an intuitive difference between the two in terms of their deservingness of blame. How might the control theorist explain this difference?

One thought is that intuitive responses to these two addicts turn on a vagueness in what sort of control an addict lacks. There are many ways an ability to resist an errant motivating desire can be compromised, and being vague about these possibilities can leave to intuitive confusion. As stipulated in the cases, the two unwilling addicts are compelled to act on their addictive desires to use drugs *once they have those desires*. Borrowing a distinction from Al Mele (1990), we might say that the addicts are unable to directly *conquer* their motivating desires head-on when they experience them. Yet, what complicates the cases is that it's left unclear whether the aspirational addict could succeed if she tries different tactics like the original unwilling addict. For instance, even if addictive desires are unconquerable once incited, she might be able to *circumvent* desires before they even arise, thereby preventing the need to try to conquer them.

To illustrate this distinction, consider a technique that some addicts use in their recovery. The heightened motivation involved in addiction is often described as being "cue-conditioned." The idea is that perceptions like thoughts, tastes, smells, and feelings can serve as cues that unconsciously trigger strongly

motivating desires for drugs.[119] Once triggered by a cue, an addict's ability to resist acting on the motivating desires is compromised. Some addicts work to circumvent these cues entirely, thus avoiding inciting the addictive desires and having to try to conquer them head-on. Through techniques like avoiding triggering friends and situations, attending cognitive behavioral therapy or group meetings to learn to redirect thoughts and feelings, and even taking medications to reduce vulnerability to unavoidable cues, addicts have, with varying degrees of success, circumvented addictive motivating desires even if they have a compromised ability to conquer the addictive desire itself.[120]

This example draws out the complications with using hypothetical cases that stipulate an inability to resist using drugs without further explaining the sort of inability involved. Returning to the two unwilling addicts, as stipulated they seem lack the ability to *conquer* their addictive desires, and in this sense they are compelled to use drugs. By exhausting all options, the original addict also clearly displays an inability to *circumvent* his errant motivating desires as well. In contrast, the aspirational addict's case requires more elaboration to appreciate what her lack of effort means in terms of her abilities. As with the resigned addict, there might be good reasons for the aspirational addict to be demotivated to try different strategies to resist her errant desires. One plausible reason would be a compromised ability to reason through the options that could actually help her resist her desire to use drugs. Alternatively, it could be the case that the aspirational addict has the ability to circumvent her addictive motivating desires and so she is less worthy of an excuse for her wrongdoing. The distinction between conquering and circumventing a motivating desires likewise helps explain why the akratic actor is more blameworthy than the original unwilling addict. Even if she cannot conquer her errant motivating desires, the akratic actor does possess the ability to circumvent them.

IV. Conclusion

The more immediate point of my argument in this paper is that the importance of concerted effort for ascriptions of blameworthiness can be explained by both mesh theories and control theories, but only the latter can plausibly explain ascriptions for cases in which effort might be weak or absent. The broader upshot of my argument is that theories of responsibility that appeal to abilities as the primary consideration for ascribing blameworthiness have more explanatory robustness than theories that appeal to identification. Of course, I've only considered the unwilling addict here and much more needs to be said about the willing addict as well since the contrast between the two addicts is what's typically meant to show the explanatory power of an agent's identification with or against her motivating desires. However, hopefully I've successfully shown that there are important differences in how addicts might lose control, and clarifying these differences in the description of cases of addiction can impact what they reveal about the conditions of moral responsibility.

Bibliography

Arpaly, N., & Schroeder, T. (2014) *In Praise of Desire*. Oxford University Press.

Capps, B., Hall, W., and Carter, A. "Addiction" in *Encyclopedia of Applied Ethics 2nd Edition.* Elsevier: p22-30.

Carter, A., & Hall, W. (2013) "Ethical Implications of Research on Craving" in *Addictive Behaviors*, 38(2): p1

Fischer, J.M. (1987) "Responsiveness and Moral Responsibility" in Schoeman, F. (Ed.) *Responsibility, character, and the emotions/ new essays in moral psychology*. Cambridge University Press:p 81-106.

[119] See Carter and Hall (2013) and Robinson and Berridge (2003). Levy (2012, 2014) helpfully details the science involved.

[120] McConnell (2016) and Kennett and McConnell (2013), for instance, argue that changing one's self-narrative is highly effective in supporting an addict's intentional planning in the process of recovery.

Fischer, J. M., & Ravizza, M. (2000) Responsibility and Control: A Theory of Moral Responsibility. Cambridge University Press.

Fischer, J. M. (2012) "Semicompatibilism and its rivals" in *The Journal of ethics*, Vol. 16 No. 2: p117-143.

Flanagan, Owen. (2017) "Willing Addicts? Drinkers, Dandies, Druggies, and Other Dionysians" in *Addiction and Choice: Rethinking the Relationship* eds. Nick Heather & Gabriel Segal. *Oxford University Press:* p66-81

Frankfurt, Harry. (1971) "Freedom of the Will and the Concept of a Person" in The Journal of Philosophy. Vol. 68 No. 1: p5-20.

Husak, Douglas (1999) "Addiction and Criminal Liability." *Law and philosophy* 18.6: p655-684.

Husak, Douglas (2000) " Liberal Neutrality, Autonomy, and Drug Prohibitions" in *Philosophy & Public Affairs* Vol. 29, No. 1: p43-80

Husak, D. N. (2004) "The Moral Relevance of Addiction" in *Substance Use & Misuse* Vol. 39 No. 3: p399-436.

Jaworska, A. (2016) "Identificationist Views" in *The Routledge Companion to Free Will.* Routledge: p37-48.

Kennett, J., & McConnell, D. (2013) "Explaining Addiction: How far does the reward account of motivation take us?" in *Inquiry* Vol. 56 No. 5: p470-489.

Leon, M. (2001) "The Willing Addict- Actor or (Helpless) Bystander? *Philosophia*, Vol. 28 No.1-4: p437-443.

Levy, N. (2006) "Autonomy and addiction" in *Canadian Journal of Philosophy* Vol. 36 No. 3: p427-447

Levy, N. (2012) "Autonomy, responsibility and the oscillation of preference" in *Addiction Neuroethics: The Ethics of Addiction Neuroscience Research and Treatment* eds. Adrian Carter, Wayne Hall, and Judy Illes. Academic Press: p139-151.

Levy (2014) "Addiction as a Disorder of Belief" in *Biology and Philosophy* Vol. 29 No. 3: p337-355.

McConnell, D. (2016) "Narrative self-constitution and Recovery from Addiction" in *American Philosophical Quarterly* Vol. 53 No. 3: p307-322.

Mele, Al. (2004) "Action, Volitional Disorder and Addiction" in *The Philosophy of Psychiatry: a Companion* ed. Jennifer Radden. Oxford University Press.

McKenna, M., & Pereboom, D. (2016) Free will: A Contemporary Introduction. Routledge: p207-216.

McKenna, M., & Van Schoelandt, C. (2015) "Crossing a Mesh Theory with a Reasons-Responsive Theory: Unholy Spawn of an Impending Apocalypse or Love Child of a New Dawn?" in Agency, Freedom, and Moral Responsibility eds. Andrei Buckareff, Carlos Moya, Sergi Rosell. Palgrave Macmillan: p44-64.

Morse, S. J. (2017) "Addiction, Choice and Criminal Law" in *Addiction and Choice: Rethinking the Relationship* eds. N. Heather and G. Segal. Oxford University Press: p426-445.

Percy, Jennifer. (10 Oct 2018) Trapped by the 'Walmart of Heroin'" in The New York Times Magazine. https://www.nytimes.com/2018/10/10/magazine/kensington-heroin-opioid-philadelphia.html. Accessed 10 Oct. 2018.

Pickard, H. (2015) "Psychopathology and the Ability to Do Otherwise" in *Philosophy and Phenomenological Research* Vol. 90 No. 1: p135-163.

Sripada, C. (2015) "Moral Responsibility, Reasons, and the Self" in *Oxford Studies in Agency and Responsibility* Vol. 3: p242-264

Sripada, C. (2016) “Self-expression: A deep self theory of moral responsibility” in *Philosophical Studies* Vol. 173 No.5: p1203-1232.

Sripada, C. (2017) “Frankfurt’s Unwilling and Willing Addicts” in *Mind* Vol. 126 No. 503: p 781–815.

Talbert, M. (2008) “Implanted desires, self-formation and blame” in Journal of Ethics & Social Philosophy Vol. 3 No.2: p1-18.

Talbert, M. (2016) *Moral Responsibility: An Introduction*. John Wiley & Sons.

Watson, G. (1975) “Free Agency” in *The Journal of Philosophy:* p205-220.

Watson, G. (2004). “Two Faces of Responsibility” in *Agency and Answerability: Selected Essays*. Oxford University Press.

Climate Change and Quality of Life[121]

Peter Railton

Abstract

Psychological research on human happiness has developed a vast literature on what is called *subjective well-being*, which is a measure that combines positive feelings and a sense of how well one's life is going. This research spans many decades, making it possible to trace the evolution of measures of subjective well-being, and also reaches to countries around the world. The research has yielded both impressive regularities and a number of notable paradoxes. Might a philosophical perspective enable us to bring together the regularities and the paradoxes under a unified understanding that tells us something important about human happiness? And might this in turn help us see why the steps needed to address long-term global climate change need not pose as stark a trade-off with short-term happiness as many have thought? I will argue that both of these questions can receive an affirmative answer.

Introduction

Bringing about the changes requisite to significantly slowing or reducing the negative effects of human activity on climate and biodiversity will require widespread changes in the ways we live. The need for these changes is dire, and is accompanied by serious questions of injustice as well—not only to future generations, but also to vast numbers of people alive today in less developed countries, who are already suffering many of the most severe effects of the failures of the more developed world to take necessary steps, and who have the least resources to contend with these challenges. A major obstacle to generating sufficient support for such changes in the developed world is the widespread belief that the requisite changes in the ways we live would come at the expense of our personal and social well-being. Growing awareness of the magnitude of the changes needed may even be lending greater force to this thought.

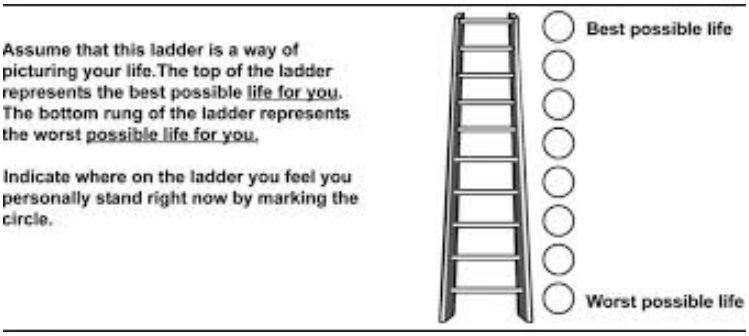
But what if this thought is misguided, not simply on empirical grounds, but in a way that reflects a persistent yet fundamental misunderstanding of the nature of well-being? Here I don't have in mind as the relevant contrast a philosophical ideal of well-being, but rather well-being of the kind that figures in the ordinary discourse and lived experience of everyday life. We in the highly-developed world might are already sacrificing our happiness, daily. How unfortunate, then, that the sacrifice might be doing more on the whole to contribute to the world's environmental problems and injustices than to counteract them.

To begin to answer this question, we must look more carefully into what is known about human well-being and its sources. This will require us to attempt to understand some of the implications of the large literature on "subjective well-being" in psychology, and to situate subjective well-being within a more general theory of the nature and fittingness of affective attitudes. Once this is done, we can return to the questions of global climate change and environmental injustice.

[121] *Preliminary Draft*, please do not quote or circulate without permission

Subjective well-being

To start our brief investigation of the psychological literature on "subjective well-being", let's look first at the elements of this measure. Although techniques vary somewhat, by far the most common way of measuring subjective well-being is via self-reported answers to survey-like questions. These questions fall into two categories: questions concerning overall life satisfaction and questions concerning current or recent mood or "hedonic tone". The scale used for questions of life-satisfaction is typically a "Cantrill Ladder", a "self-anchoring" score in which individuals are encouraged to give a response placing them at one or another "rung" on a ladder running from the best or worst possible life for themselves:



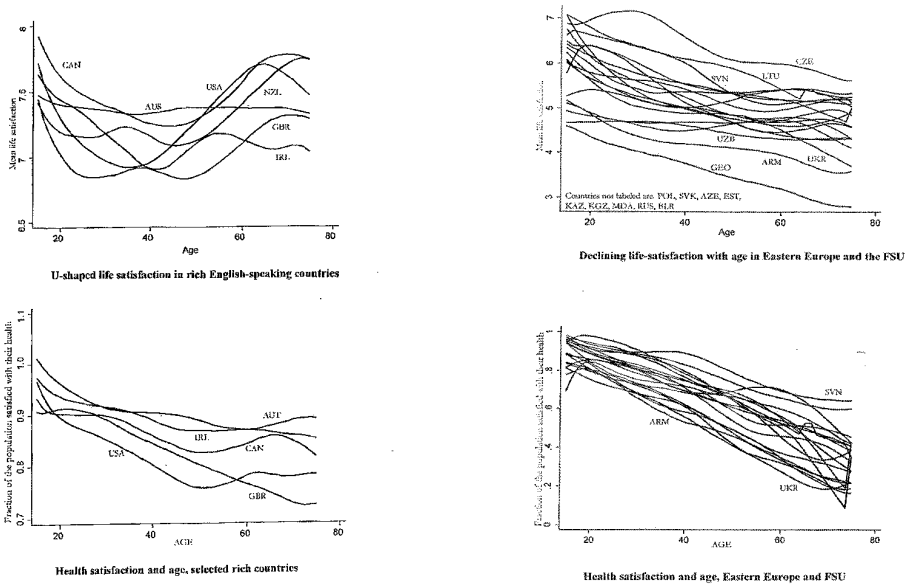
Or the rungs of the ladder might stretch from "Very satisfied with my life as a whole" to "Very unsatisfied with my life as a whole", or from "Thriving" to "Suffering". The scale used for the second, affective component of subjective well-being is typically a "Likert Scale", which runs (often horizontally) as a continuous line from negative to positive, and where individuals are asked to place themselves along the line in terms of how strongly or frequently they have felt various kinds of affect, e.g., good mood, blue mood, stressed, lonely, etc., over the past few days. The value of the two measurements is then taken together, assigning each equal weight, and taking the average, yielding "subjective well-being". In addition to surveys, subjective well-being is measured in interviews, via periodic random sampling via cell-phone, and so on.

These are self-report measurements, and one might expect them to show a large amount of arbitrary variability as a result. They *do* in fact show variability, depending upon various factors, including the particular context or framing of the sampling, but much of this variability is systematic, and can be taken into account in assessing the results in a given sample (for extended discussions, see Diener, Schwartz, and Kahneman, 1999). Moreover, while there are many methodological issues about subjective well-being, as a measurement it is statistically fairly well-behaved and shows a number of stable relationships to more "objective" measures of a person's condition—whether demographic, social, financial, or physical. And these measures also turn out to have predictive value—knowing someone's subjective well-being enables one to better predict various social or health outcomes than relying upon "objective" measures alone. As a result, measuring subjective well-being has gained widespread acceptance, and a large industry has grown up around the various uses of subjective well-being in research, clinical practice, and public policy. It also turns out that the public has a large appetite for this kind of measurement, and governments are beginning to use assessments of subjective well-being as a leading social indicator, in addition to such indicators as economic growth and employment.

Philosophers have tended to be skeptical about subjective well-being, especially given its reliance upon self-report and the ways in which measurements of subjective well-being are so often presented in the press as measures of *happiness* or *well-being*, full stop. Psychologists have not all been suitably cautious on this score, but they often are careful to emphasize that subjective well-being is not an attempt to capture the whole of human happiness, and make apologies in the direction of philosophy for using a measure of well-being seems to fall so far short of historically important philosophical conceptions of happiness or well-being such as *eudaimonia*. In return, however, philosophers probably owe it to psychologists, and to humankind more generally, to attend to what the psychologists' results in measuring subjective well-being might be telling us about a central element in the intrinsic value of a life, namely, how that life is experienced or appraised by the one leading it. A philosophical account of personal well-being cannot afford to be so far alienated from lived experience as to treat as irrelevant what people themselves will say when asked to reflect on how well their lives are going, or how they feel day-in and day-out.

Attention to the literature on subjective well-being may also help philosophers keep some aspects of a normative theory of well-being in proportion. For example, in some countries, self-reports of how well life is going or how one feels daily decline markedly in the final decades of life, while in others it actually rises (the so-called "U-shaped curve" of subjective well-being). This is so even though self-evaluated *health* is almost universally self-reported as being in decline during these decades. For instance:

Fig. 1. Satisfaction with life vs. satisfaction with health as a function of age: wealthy Anglophone vs. former Soviet Bloc countries



Thus, while decline in health in the final decades of life is virtually inevitable—as well as related losses in terms of mobility, memory, perceptual acuteness, and ability to manage daily life independently—it does not appear to be inherent in the human condition that this should reduce the experienced value of one's life. While psychologists interested in happiness should not assume that subjective well-being constitutes *eudaimonia*, philosophers interested in happiness should not assume that erosion, absence, or loss of the "full development of species-typical human potential" necessarily tears at the foundations of human well-being. One of the early results of research on subjective well-being was that individuals who have suffered a disabling accident return after a year or two to a relatively high level of self-reported well-being (Brickman *et al.*, 1978). More generally, disabled individuals and disability advocates have argued forcefully that the easy assumption that full human flourishing requires full development of human potential misrepresents the lived experience of disabled individuals, and leads to a failure to learn about the sources and nature of human well-being from these lives (cf., for example, Solomon, ref.).

The evidence represented by Fig. 1 should help strengthen this argument—theorizing about well-being should be informed by investigation into the factors explain the differences in experienced quality of life in life's final decades of life across these different countries and cultures. For example, the extreme decline in support for the elderly in post-Soviet Eastern Europe meant that deteriorating health and a deteriorating sense of well-being went hand-in-hand. But societies that provide higher levels of support for the elderly make it evident that this can be one of the most satisfying periods of a person's life, despite substantial loss of certain dimensions of human potential. Moreover, the contrast between Eastern Europe and Anglophone countries helps establish that "adaptive preferences" ("settling for less") cannot be the main story behind this period of relatively high life-satisfaction—that mechanism would not predict the steep decline in subjective well-being in Eastern Europe.

But haven't psychologists themselves thrown doubt upon "intuitive" judgments of this kind, as the product of "System 1" processing and thus given to simple heuristics, and fast, affective rather than cognitive responses (Haidt, 2001)? And indeed psychologists have documented many ways in which people lack insight into their own psyches (Nisbett & Wilson, ref.) and their responses on measures of subjective well-being seem sensitive to seemingly irrelevant or very transient features of the context in which the questions are asked or the framing used in asking them. It is not difficult to imagine that cultural and ideological differences in common notions of what makes a life a good one will make large-scale or cross-cultural comparisons fairly meaningless. And well-known individual tendencies toward self-validation and overly favorable self-assessment and comparison with others (some three-quarters of drivers rate themselves as "above average"), seems likely to skew answers toward the positive in a way that doesn't reflect any underlying reality in the quality of life. Why then should we be paying attention to self-reported affect, mood, or life-satisfaction as a ground for serious discussions about human well-being, much less as a benchmark for assessing policies or polities?

These are all legitimate worries, though some of them, I believe, rest on dubious assumptions (see, e.g., Railton, 2014). Moreover, since psychologists working on subjective well-being have been the source of much of this critical research, they have devoted considerable effort to studying contextual and systematic effects on reported subjective well-being, seeking to determine whether or when it has predictive value, and applying technical statistical measures the data generated. As a result, many of the first- and second-generation criticisms of measures of subjective well-being have been addressed or mitigated. While many problems remain, I am less skeptical than most philosophers about measures of subjective well-being and their results. Such measures do, I believe, convey genuinely important information about the quality of people's lives, though perhaps not quite the information their most ardent partisans think.

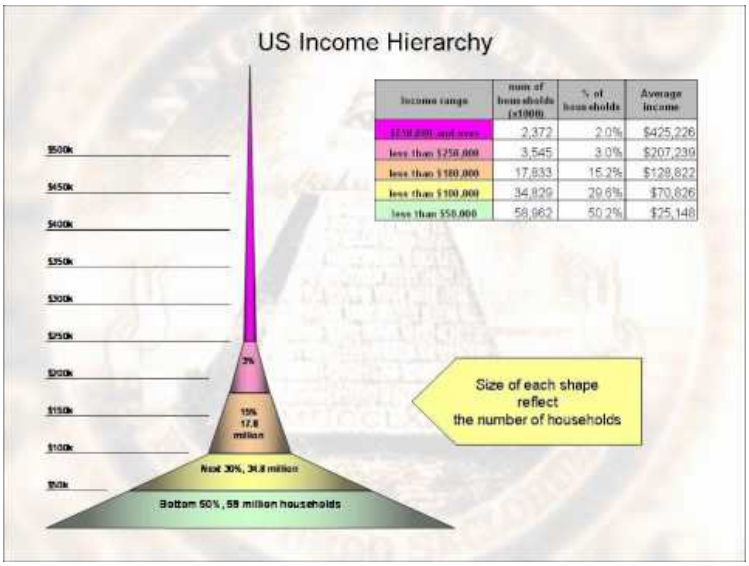
Too little variability?

Standardized measures of subjective well-being began to be used systematically in World War II, trying to understand how well individuals bear up in, or after, combat situations. Once the war was over, the measures began to be used to observe trends in the population more broadly.

A first surprise was that the overall figures *weren't* very trendy. Variability in individual reports was not as great as one might have expected—the great bulk of the population in countries like the US and the UK crowded into the top half of the scales, with an average of 3.8-3.9 on a five-point scale not uncommon. And distribution around this mean was "normal", not bimodal, falling off on either side of the mean in ways typical of well-behaved variables, resulting in a fairly thin tail by the time one reaches the bottom of the range. So the modal subjective well-being measured in well-developed countries isn't "neutral" (mid-way on a scale, say), and individual fluctuation over time tends to stay within a range centered on the modal value, spending some time above, and some time below, but in a fashion that is fairly balanced within a wide sample. And as we'll see, within a country the modal value tended to remain fairly stable over time, despite significant changes in other variables.

One significant way in which there was less variability than expected tended to confirm nineteenth-century Utilitarian theory: income shows a highly reliable diminishing marginal "utility".

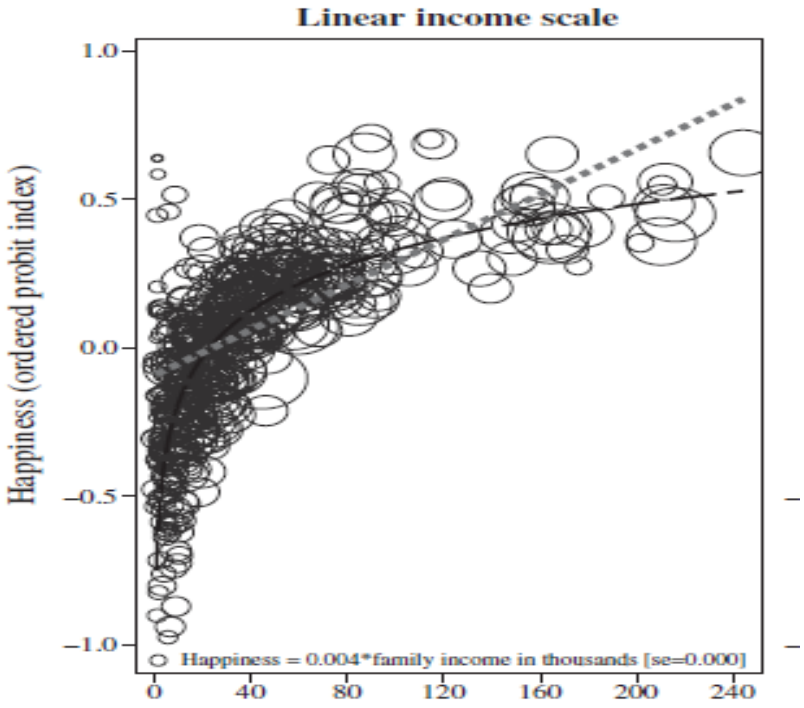
Above a certain level of income, which varied from country to country, but which typically corresponded to a fairly middle-class level of income for that country—increases in income had rapidly diminishing marginal positive effects on people's subjective well-being. In contrast to the official skepticism among economists about such comparisons, both across individual lives and within individual lives over time, to listen to people tell it, gains or losses in income matter much more if they occur to those with few financial or social resources. Note first the shape of the income hierarchy in the US in 2009:



Income range	num of households (x1000)	% of households	Average income
$250,000 and over	2,372	2.0%	$425,226
less than $250,000	3,545	3.0%	$207,239
less than $100,000	17,833	15.2%	$129,822
less than $100,000	34,829	29.6%	$70,826
less than $50,000	58,962	50.2%	$25,148

Compare this with the shape of the curve for subjective well-being vs. income:

Fig. 2a. Subjective well-being vs. income: US 1972-2006 (General Social Survey)



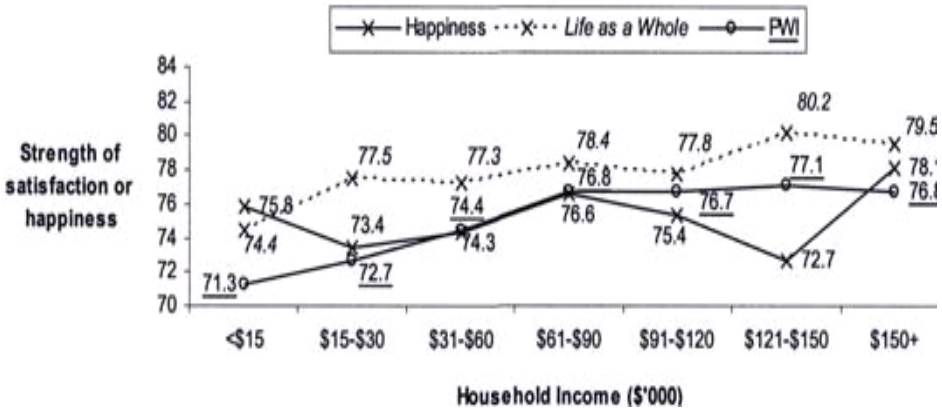
Source: General Social Survey.

Consistently with the idea that a reasonable level of *sufficiency* in income plays a larger role in people's subjective well-being than the absolute level of income, the pattern of growth in subjective well-being in relation to income in countries where there are greater *social* resources for those in the lowest quintile, shows a shallower rate of return from income:

Table 1. Subjective well-being vs. income: Switzerland (ten-point scale) (Leu, Burri, and Priester, 1997)

- Highest income 1/5th 8.45
- Next highest 8.49
- Middle 8.24
- Next lowest 8.17
- Lowest 7.98

Fig. 2b. Expressed happiness and life satisfaction vs. income: Australia 2006



Despite the intense focus on income as a gauge of well-being, it appeared that, once a level of reasonable sufficiency was met, either through actual income or through social provision of resources, income increases added surprisingly little to subjective well-being. Of course, these statistics do no break out *extremely high income* individuals, earning millions a year. Perhaps the curve inflects back up for the super-rich? Anyone who has known super-rich individuals is likely to doubt that they as far above the average happiness of the upper 5% in income as they are above the average income in the upper 5%.

A similar lack of variability in reports of subjective well-being showed up in cross-personal longitudinal data. Here, for example, are the curves for real income (corrected GDP per capita) vs. subjective well-being for the US in the period 1945-1991:

Fig. 3. Real GDP per capita vs. subjective well-being: US, 1946-1991

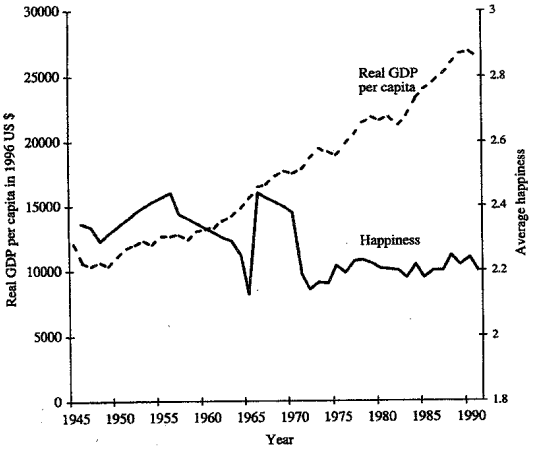


Figure 4.1. Happiness and income per capita in the United States, 1946–91. Data from World Database of Happiness, Bureau of Economic Analysis of the U.S. Department of Commerce and U.S. Bureau of the Census.

Apart from some interesting activity in the 1960's, Fig. 3 shows little net change in average reported "happiness"—and certainly no net *growth* despite a very considerable increase in the material standard of living. Extending the curve to 2003 does not reveal any striking reversal of this pattern:

Fig 4. Subjective well-being vs. real per capita income: US, 1973-2004

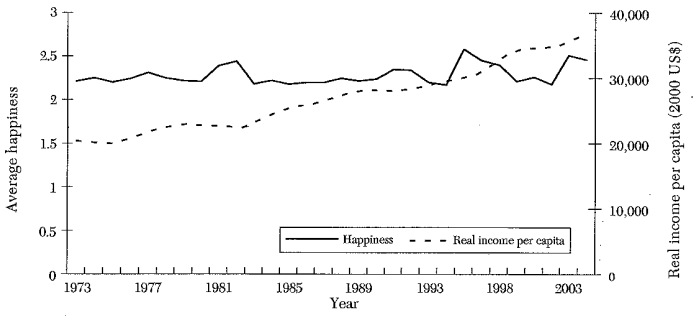


Figure 1. Happiness and Real Income Per Capita in the United States, 1973–2004

Source: World Database of Happiness and Penn World Tables. Happiness is the average reply to the following question: *"Taken all together, how would you say things are these days? Would you say that you are…?"* The responses are coded as (3) Very Happy, (2) Pretty Happy, and (1) Not too Happy. Happiness data are drawn from the General Social Survey.

One might fish for explanations in the peculiar political and cultural trajectory of the US, except that a similar phenomenon has occurred many countries in Western Europe, where the real material standard of living has risen even more dramatically since the 1960's than in the US. Here is one long-running European survey:

Fig. 5. Percent "Very Satisfied" with life as a whole: Europe, 1973-1998

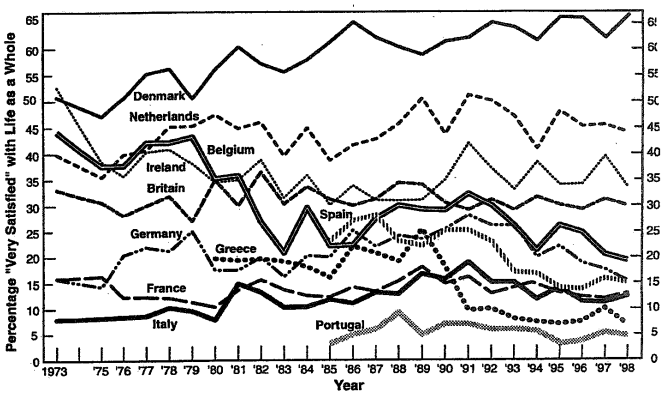


Figure 7.1
Cross-national differences in satisfaction with one's life as a whole, 1973–1998. Source: Euro-Barometer surveys carried out in each respective year.

Indeed, only the world-beating Danes show a noticeable upward trend over this period. If we confine our attention to the UK, Germany, France, Netherlands, and Italy, the lack of change in average satisfaction with life is yet more obvious:

Fig. 6. Average satisfaction with life as whole: 5 European countries, 1973-2004

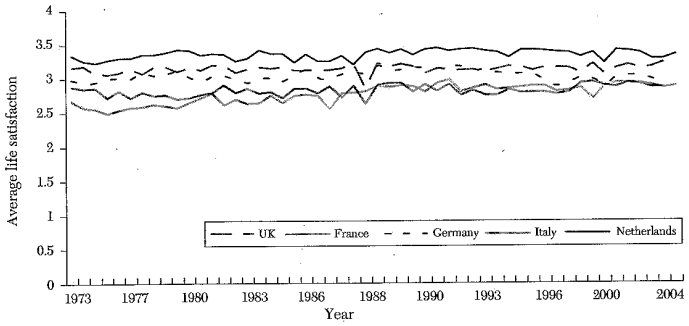


Figure 2. Life Satisfaction in Five European Countries, 1973–2004

Source: World Database of Happiness. Happiness is the average reply to the following question: *"On the whole how satisfied are you with the life you lead?"* The responses are coded as (4) Very Satisfied, (3) Fairly Satisfied, (2) Not Very Satisfied, and (1) Not at all Satisfied. Life satisfaction data are drawn from the Eurobarometer Survey.

This, despite the fact that gains in real per capita GDP ranged from 70% in France (from $13,700 in 1974 to $22,600 in 2004) to a doubling in the UK (from $13,800 to $27,800). It would appear that the creation of a staggering about of material wealth within these countries since the 1970's has not had the effect of substantially changing average subjective well-being at all.[122] Indeed, the flatness of the curve in (Fig. 5), despite the economic and political ups and downs of these countries during the last three decades—to say nothing of the significant social and demographic changes they underwent in the years that followed the

[122] Throughout, unless otherwise noted, I will follow the standard usage in the social science literature of speaking of "statistical effects", such as regression coefficients, as "effects" *tout court*. This, I grant, is tendentious and potentially quite misleading, since "effect" is also a *causal* term, and very much less is known about causal effects upon, or of, subjective well-being.

1960's—might lead one to think that average subjective well-being simply isn't an *informative* social indicator after all. (We will see below some very important qualifications to this idea.)

Suppose that one extends this research beyond the boundaries of the most developed countries. Do the data show the dramatic effects from income that one might have expected? Looking at a wide swath of nations for the period 1980-1995, one does see a significant overall upward trend in average subjective well-being as a function of per capita "estimated purchasing power":

Fig. 7. Subjective well-being vs. per capita GDP across countries: 1981-1995

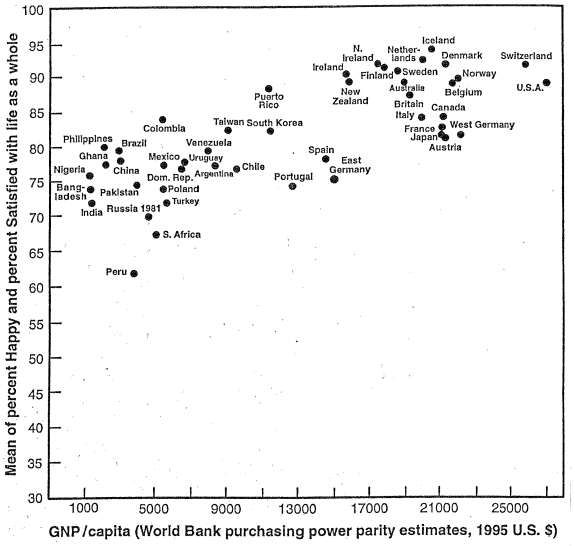
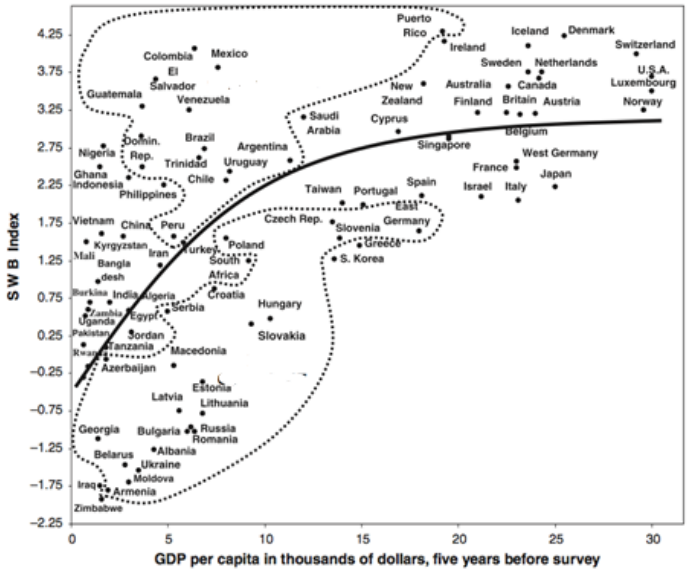


Figure 7.4
Collapse of communism and the decline of subjective well-being in Russia. The correlation between GDP/capita and subjective well-being, omitting the former communist societies, is $r = 0.74$. Source: World Values Surveys. Data for Russia in 1981 from Tambov oblast, 1981.

But even so, many polities with half or even one third the average per capita purchasing power of the richest countries nonetheless enjoy equal or higher average subjective well-being—Puerto Rico, Colombia, and Taiwan are at or above the level of France, West Germany, or Japan. Moreover, above $17,000 in average per capita purchasing power, there is significant variation, but it has no association with rising income, despite thousands of dollars of increase in GDP per capita. It appears that, in 1981, once purchasing power parity GDP per capita reaches a level of approximately $17,000, other factors account for the scattering of countries above or below the trend-line in Fig. 8. Even if we look further much down the income scale, $7,000 or less in 1981, we find that the percentage of respondents in Ghana, India, Bangladesh, Poland, Mexico, and Turkey reporting they are "happy" and "satisfied with life as a whole" drops only to 70%. Moreover, it appears that there are important social and cultural factors underlying much of the variation, not dependent upon income:

Fig. 8. Subjective well-being vs. per capita GDP across countries, ca. 2008



Source: Ingelhart *et al.*, 2008

Recently this picture of the relationship between income and subjective well-being, which had achieved fairly wide acceptance in the scholarly literature—even earning the name, 'the Easterlin Paradox' in honor of an economist who gave it prominence, has been challenged by other economists. Stevenson and Wolfers (2008, 2013) point out that decreasing marginal utility as a function of raw absolute income is compatible with *increasing* total utility as a function of *log*(absolute income), and argue that recent global Gallup polling in fact bears this relationship out.

However, the life satisfaction portion of the Gallup poll is a "ladder" poll, and this design is known to encourage individuals to think of their estimated position on the socio-economic "ladder" in their society, judging their relative social standing rather than directly reporting their experienced sense of satisfaction with life (see also Diener *et al.*, 2013). Moreover, even so the *rate* of increase in ladder score fell off with increasing *log*(income) as well as increasing income. If we break the subjective well-being score into its components in the Gallup poll, it becomes clear that the small increase in reported subjective well-being with *log*(income) does not reflect a continued increase in positive affect. Thus Kahneman and Deaton (2010, see also Diener *et al.*, 2010) found:

Fig. 10. Positive affect, negative affect, stress, and life evaluation vs. *log*(income): US 2009-2009 (Kahneman & Deaton, 2010)

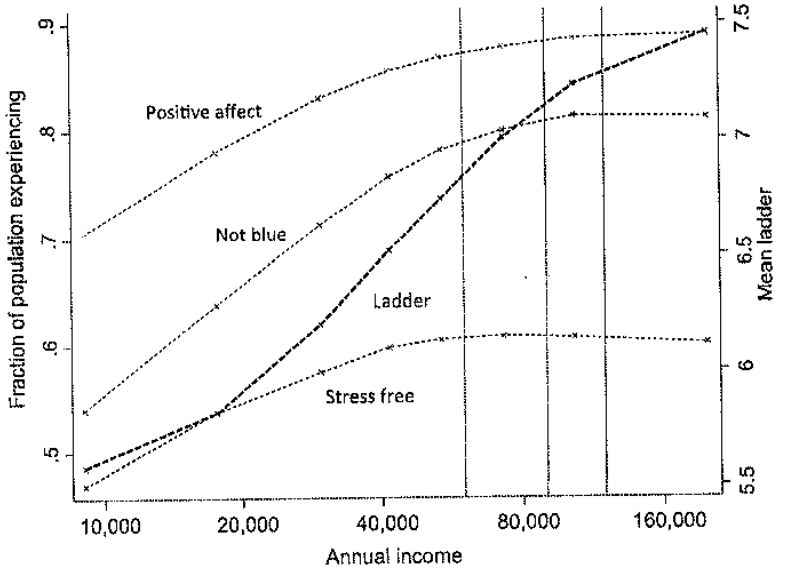


Fig. 1. Positive affect, blue affect, stress, and life evaluation in relation to household income. Positive affect is the average of the fractions of the population reporting happiness, smiling, and enjoyment. "Not blue" is 1 minus the average of the fractions of the population reporting worry and sadness. "Stress free" is the fraction of the population who did not report stress for the previous day. These three hedonic measures are marked on the left-hand scale. The ladder is the average reported number on a scale of 0–10, marked on the right-hand scale.

Fig. 10 indicates that *log*(income) does indeed yield continuing increases in overall life satisfaction well beyond the point of reasonable sufficiency (circa $85,000), though it is worth noting that the scale in Fig. 10 extends only as far as approximately $240,000 in annual earnings, and that this curve, too, becomes concave above approximately $90,000. This suggests that life satisfaction shows diminishing marginal returns against *log*(income) as well as income. And even taking *log*(income) rather than income as a metric, positive affect and freedom from negative affect and stress show clearly diminishing marginal returns from income above $40,000, and essentially flatten above approximately $85,000.

Thus, while it is true that the Gallup poll supports the idea that at least one component of subjective well-being, overall life satisfaction as measured on a "ladder" scale, may continue to rise as real income rises, the relationship is logarithmic, so that an *exponential* increase in income is needed to make each successive—increasingly small—incremental gain above $90,000. An income in the hundreds of thousands of dollars might have to be *doubled* to achieve a barely significant effect. From the standpoint of social policy or personal decision-making, the larger lesson is that this would seem to be a highly resource-intensive way of securing very little gain in subjective well-being.

Easterlin's own most recent work, especially on China, along with related work on India (Deaton, ref.), points to the conclusion that raising per capita GDP alone cannot be expected to raise subjective

well-being across a wide population or even, in the case of India, to improve the meeting of caloric needs. By contrast, policies that provide for "full employment and a generous and comprehensive social safety net" do make a reliable contribution to subjective well-being, even in relatively poor countries (Easterlin, 2013). Since many policies oriented toward rapid economic growth have a tendency to displace population in ways that can increase unemployment and precariousness, without instituting a replacement set of social institutions (e.g., in China and India), raising GDP per capita in such societies has not tended to produce significant gains in average levels of subjective well-being.

"Adaptation-level theory": Comparison, habituation, and the "hedonic treadmill"

A fairly standard explanation of the relatively weak marginal effects of higher income on subjective well-being, and the surprising stability of national levels of subjective well-being in developed countries despite very substantial real economic growth (e.g., Figs. 5-6, above), is that that people *adapt* to such changes, raising their expectations, and thus show no net long-term gain in subjective well-being. Indeed, according *adaptation-level theory* (Brickman & Campbell, 1971, Brickman *et al*., 1978), two distinct processes operate conjointly to have the effect that a rise in income or wealth, while producing some gain to subjective well-being ("happiness is relative to a reference point"), these effects will tend to attenuate over time ("people habituate to their environment"), leaving things essentially unchanged. The long-term level of subjective well-being thus tends to reflect, not economic gains or losses, but a relatively stable distribution in society of underlying personal dispositions toward a positive experience, sometimes called individual "set points". Individual dispositions themselves tend to be fairly stable over time, and related to basic personality variables. Let us look briefly at these processes, and how they work together to produce "hedonic adaptation".

Happiness is relative to a reference point. The question, "How well is your life going?", might naturally lead to another: "Compared to what?", but according to *social comparison theory* people already have an answer to that second question in the form of implicit standards of comparison used in everyday life. Such standards might be given in a particular context by co-workers, neighbors, kin, etc., and different answers to the question "How well is your life going?" can be primed by priming different comparison classes.

Someone who learns that she will receive a 15% raise will tend to compare this with her co-workers, and, should the comparison be favorable, will tend to experience a gain in positive affect and life-satisfaction. This gain will typically occur before the increased income has made any material effect upon her life. That should not be puzzling—one very common goal in life is to "do well", and a 15% gain in income relative to co-workers constitutes significant evidence that one is "doing well". However, should she at the same time learn that her co-workers are receiving large raises, this would be a weaker sign of "doing well", and the change in her subjective well-being is likely to less than the material improvement made possible by the raise would predict. Luttmer (2005) found, for example, that, for a given absolute level of income, those whose neighbors had higher earnings than themselves had lower self-reported happiness than those who earned more than their neighbors. Assadullah *et al.* (2018) found a similar effect in China, especially for males. These responses to relative rather than absolute position can be quite finely tuned and contextually-sensitive. In an early experiment, Brickman (1975) found that, for a small benefit of fixed size, individuals in an experimental group were more satisfied with the benefit

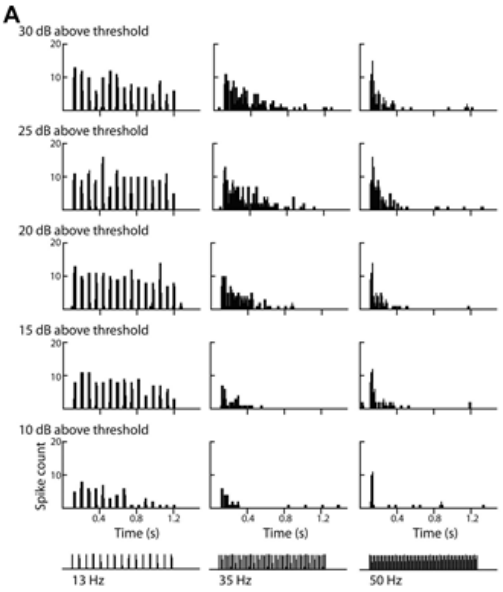
when one member of the group did not receive it than if everyone in the group received the benefit, and less satisfied if one member of the group received more. Thus, during years of widespread economic growth, gaining income while others are doing so as well could be expected to have a mitigated effect overall, at least once one is above a level of sufficiency in income. Moreover, each time someone gains *substantially* more than others, these others will suffer somewhat by comparison. So *net* gain in average subjective well-being from generalized, incremental, but unequally-distributed or temporally-staggered increases in real income could be much reduced.

Moreover, comparison effects can work against stable high gains from income even when the gain is very large. A classic study by Brickman *et al.* (1978) found that even those who had won a major lottery within the last 6-18 months (for some, the prize was $1 million in 1978 dollars), were not nearly so far above the population average as one might have expected—or they might have imagined beforehand. On average, they measured 4.0 on a 5-point scale of "general happiness", as against 3.82 for controls who had experienced no such windfall. Winning a large sum certainly moved the individual "ahead" compared to his previous level of wealth, but at the same time this highly salient *large* piece of exceptional good fortune made other, everyday sources of subjective well-being—time spent with friends, leisure time with family, watching television, etc.—seem comparatively lackluster. This predicts that lottery winners should "get less" from these everyday sources of pleasure or satisfaction, and indeed Brickman *et al.* found that lottery winners reported a level of enjoyment of everyday activities of only 3.33, compared to 3.82 for controls. Thus the gain in subjective well-being from comparative advancement in wealth was to some extent intra-personally offset by a loss from comparative diminution of magnitude of mundane pleasures.

But now we face a seeming puzzle: It is one thing to find that comparison effects can *mute* long-term gain in individual and social subjective well-being, but another thing to explain social trends like the stability of average reported subjective well-being the UK, France, Germany, Italy, and the Netherlands during the years 1973-2004, despite large but differing levels of growth in per capita GDP during this period and varying social and political histories.

Individuals habituate to their environment. To answer, Brickman *et al.* (1978) posited a middle- to long-term mechanism borrowed from experimental psychology, *habituation.* In habituation, a novel stimulus that initially excites a response will lose its ability to evoke any noticeable response if it is simply repeated without variation. Upon entering a room with a ticking clock or buzzing light fixture, the noise will annoy at first, but as one's mind turns to other things the sound will disappear from audition, and only a deliberate shift of attention, or a variation in the stimulus, will reinstate the sound to audibility.

Fig. 11. Auditory midbrain response to a repeated sound (Elliott *et al.*, 2011)



The advantage of this feature of sensation is two-fold. First, if a stimulus actually shows no variation, then it is no longer a source of new information about the environment, so attention and auditory bandwidth need not be devoted to it. Second, this self-attentuating effect leaves the ear more sensitive to other sounds in the environment that might carry new information. Or consider the juvenile squirrels of the spring in the Quad, tentatively descending from their nests. When a pedestrian approaches, they leap excitedly back to the tree, climbing over each other if necessary. But within a month or so the continual passing of pedestrians without incident has led them to habituate to this feature of the environment, and they pay little or no attention to it as they range around engaged in the serious business of looking for seeds or crumbs. What once produced a stressed, fearful response has ceased to have such an effect, and become so much wallpaper.

In a similar way, it was theorized, a change in lifestyle once fully routinized will no longer draw attention to itself or produce a corresponding difference to the quality of one's lived experience. Whether one is living in Manchester at $13,800 in 1974 or $27,800, and whether or not one knows if one's income gains have reflected the social average, still, within one's own life one will have habituated to the income after a certain period of time. So one will tend to regress to whatever one's "normal" or "set-point" level of subjective well-being might be. Scaling up to the population as a whole, this means that income gains tend to "wash out" over a certain number of months, leaving averages pretty much where they were before the gains. The larger car or roomier house, that initially promised a long-term gain in happiness, will tend to become the new normal against which life transpires. Trying to raise a country's subjective well-being significantly simply by increasing income or material goods would be, in the long run, unlikely to succeed. As Fig. 3 suggests, a greater-than-doubling of real GDP per capita in the US from 1945 to 1995, which represents a tremendous quantity of money and goods, left average subjective well-being, if anything, a bit lower.

One might wonder, if individual subjective well-being inevitably regresses back to a personal "set point", after increases in income or wealth, what could explain the amount of energy humans devote to

gaining more income and wealth for such transient effects? Why don't they learn that they cannot really ratchet up their level of subjective well-being, at least, once a reasonable level of sufficiency has been attained?—The way one eventually learns to stop spending money on CDs or software boasting: "Speak Mandarin Like a Native: In just 45 minutes a day for two weeks!"?

One explanation is that people suffer from a persistent "front-end effect": each increase in income has brought with it some fairly strong immediate positive feelings. Even if these positive feelings attenuate, still, working for a raise *was* rewarded, and earning a new raise *will be* rewarded. The gains need not be lasting in order for behavior to be conditioned by this kind of association with reward. So most people will agree with the statement, "If only I earned a bit more, I'd be happier", although they may fail to notice the gradualness of attenuation which makes this statement more transient than lasting. By contrast, one never had a potent experience of gaining significant proficiency in Mandarin from listening to a CD, so there is no "front-end effect" to bolster the spurious promise of the advertising. So some things we learn not to waste time and money on—but money itself, we do not. The result of this habituation-cum-comparison effect has been dubbed the "hedonic treadmill", as individuals keep struggling for the immediate goal of an increase in income while their ultimate goal—achieving a higher *enduring* level of happiness—always stays tantalizingly ahead of them.

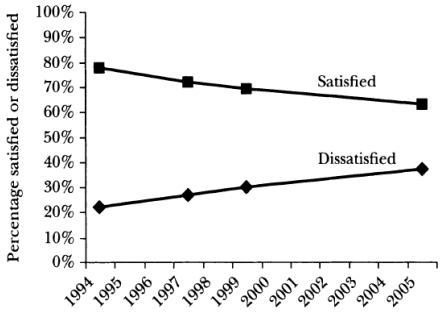
One might worry, though, that habituation *over-explains*. If people will habituate to whatever environment they find themselves in, why do we see *rising* curves from low income to moderate income within a country or across countries? Why would the "income effect" not attenuate in the same way whether one is earning $1,000, $10,000, or $25,000 per year? Income must be making some real, persistent difference to lives despite whatever habituation occurs.

And why would the typical individual in most countries have a relatively *high* rather than *neutral* reported level of happiness and life-satisfaction. In sensory adaptation, or in the adaptation of the stress level of squirrels, the return is to a neutral state, not an "above average" state. And why is the typical level high in many countries even when the economy is now growing very slowly, in comparison to the post-World War II era?

Could the other component of the adaptation-level hypothesis, that happiness is comparative, help? Could it be that the positivity comes from the way people select their reference group for comparison? By judiciously picking those *below* us, we can buoy our sense of life satisfaction. But social psychologists have found that, when people in the US are asked to name acquaintances of roughly the same economic or social standing as themselves, they typically pick a set of acquaintances 15-20% *above* them in economic or social standing. This makes sense from the standpoint of self-motivation and status aspiration (cf. Collins, 1996), but it doesn't fit with the classic picture of Brickman and Bullman (1977). A purely comparativist logic would predict that such "upward comparisons", would make most of us relatively *un*happy and *un*satisfied—like the subjects in Brickman's (1971) experiments who saw themselves disadvantaged with respect to their group. If individuals are motivated to make *downward* comparisons, as a way of strengthening self-image, this implicit strategy would seem to be readily available to contemporary Chinese, in an economy that has seen a 250% growth in real GDP per capita between 1994 and 2005. For even those who haven't gained the exceptionally large amounts can readily

identify some salient individuals who have gained less—or simply focus on the generation of their parents. But instead of the pattern this would lead us to predict we see:

Fig. 12. Percent satisfied or dissatisfied overall with life, China, 1994-2005 (as reported in Kahneman & Krueger, 2006)[123]



Source: Derived from Richard Burkholder, "Chinese Far Wealthier Than a Decade Ago—but Are They Happier?" The Gallup Organization, ⟨http://sww.gallup.com/poll/content/login.aspx?ci=14548⟩.
Notes: In 1997, 1999 and 2005, respondents were given four response categories: very dissatisfied; somewhat dissatisfied; somewhat satisfied; and very satisfied. In 1994, respondents were given a fifth response category: "neither satisfied nor dissatisfied." The chart reports the percentage who were satisfied or dissatisfied. Thirty-eight percent of respondents chose the neutral category in 1994; those respondents were allocated in proportion to the number who responded that they were satisfied or dissatisfied in that year.

Too much variation?

If subjective well-being seems to show too little variation, or the "wrong sort" of variation with respect to long-term gains in income, it has seemed to show *too much* variation with respect to some quite transient gains and losses. A study of introversion and extraversion among college students found a *stable difference* in subjective well-being at a time, but a surprising amount of *parallel variability* over time:

[123] Fig. 12 may reflect the phenomenon of growth producing rising expectations as well as the effects of significant population dislocation and loss of traditional forms of social support. More recent surveys have shown that subjective well-being has begun to recover in China, though it hasn't reached earlier levels (Easterlin *et al.*, 2013).

Fig. 13. Extraversion, introversion, and sense of well-being vs. days of the week: US college students (Larsen & Kasimatis, 1990)

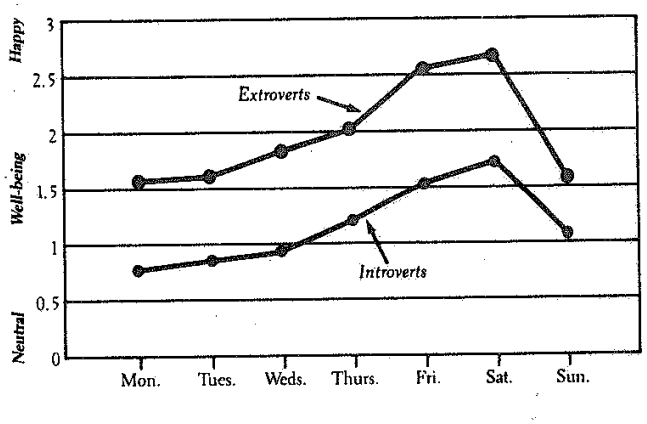


Fig. 13 suggests not only that day of the week matters to affective state, but that *anticipation* of the day to come may have a significant effect on affective state. As everyone who has gone to school or worked at a tedious job knows, Sunday, even though it is still a "free day", lies always in the shadow of Monday, hence the steep decline even before Monday arrives.

Kahneman and Krueger (2006) used day-reconstruction methods and found that experiential well-being could vary by a factor of 2 over the course of 12 hours, while a Gallup poll showed that having a headache can lower life-satisfaction more significantly than a chronic health condition (see the "ladder" value in Table 5, below). Schwarz and Strack (1999) review a number of experiments in social psychology showing how much individual responses to questions about subjective well-being varied as a function of "framing", the ordering of questions, the priming of positive or negative thoughts by trivial events, etc. could lead to significant variations in reported overall life satisfaction as well as momentary affect.

A satisfactory explanation of the stability of reported subjective well-being in the large should also give some insight into the instability of reported subjective well-being in the small. It is not clear how habituation or social comparison could accomplish this.

I think, then, that we should be looking beyond the psychic mechanisms of adaptation-level theory. We cannot rule out the possibility that there are no *systematic* psychic mechanisms at work in the sorts of self-attributions made of life satisfaction or experiential well-being—felt happiness might be a late evolutionary "add-on", perhaps a by-product of other changes in the brain with no function proper to itself. Perhaps this would explain the seemingly disunified patchwork of responses survey research has found. But before giving up, let's pursue a bit further the question whether we can identify core psychic processes that could *explain* the patterns of subjective well-being over time and across nations that the habituation and social comparison hypotheses were introduced to account for.

Habituation vs. accommodation

A first step is to distinguish habituation from two seemingly similar fundamental psychic processes with which it is often conflated—*accommodation* or *adaptation*. Habituation, as we saw, occurs when repeated exposure to an unchanging stimulus attenuates the response to that stimulus. The solar panels just installed on your roof in an effort to be ecologically-minded, and which you and your neighbors now see as so glaring, will fade in conspicuousness to you both over the course of the coming year. By year's end, you will pay them no attention at all. This, like the case of the juvenile squirrel, is an example of stress-reduction through habituation.

Stress is a sensitive indicator of degree of habituation, since stress functions to heighten vigilance, mobilize attention and cognitive resources, and prepare the organism for action—the opposite of the indifference shown to a habituated stimulus.

Therefore it is interesting that a larger, longitudinal study of winners of "moderate" lottery purses in the UK (between $1800 and $200,000) found that even two years later winners showed lowered levels of stress, and better mental health characteristics generally, than otherwise comparable controls who had not won a lottery (Gardner & Oswald, 2007). Two years after a windfall increase in wealth, individuals had not habituated back to baseline levels of reported stress or mental health. It is also interesting that public health research in the US has found that stress rises as income declines—lower-income individuals have higher levels of cortisol and other stress-related hormones in the bloodstream, with attendant corrosive effects on physical and mental health (Cohen *et al.,* 2006; Evans & Kim 2007; Li *et al.,* 2007; Lupien *et al.,* 2011; Haushofer *et al.*, 2012).

It appears that individuals do not "habituate" to the life circumstances of living on a low income in a prosperous country. Similar failure to habituate is found among the unemployed—as the duration of unemployment increases individuals experience higher, not lower, levels of stress, worrying, and negative affect (Gallup, 2010).

Low levels of income or resources makes one's life position more *precarious* even if one is, at the moment, successfully getting by—as we know, *anticipation* can significantly affect subjective well-being in the moment. Being situated thus on the margin between getting by and failing to do so, maintaining the *status quo* calls for higher levels of vigilance and economic concern. A single unfortunate event—an injury, damage to one's car or house, an expensive illness—can tip one over into failing to get by. Sustained unemployment is also such an event—depriving one of what is, for most people, their main source of income, security, and social standing. The chronically elevated levels of stress hormones found in those who are living on the margin or unemployed in prosperous countries are evidence that they are *accommodating* to their condition, not habituating: precariousness requires an adaptation for greater vigilance, economic concern, care and motivation to manage resources, gain small amounts, or preserve employment. Stress gears up the body and mind for such challenges, but when the challenged state becomes chronic the constantly elevated level of stress can lead to breakdowns in physical and mental health—and these themselves are predictors of lower subjective well-being. This is not habituation, then—the individual's system has not become oblivious to the "unchanging stimulus" of her chronically marginal economic condition or unemployment, on the contrary.

To understand how accommodation differs from habituation, an example might help. Individuals who move from low to high altitudes initially experience weakness, nausea, and difficulty breathing. Over a period of weeks, increased hemoglobin in the blood enables them to *accommodate* their physiology to the reduced level of available oxygen per unit volume of air at high altitude. Once this bodily accommodation has occurred, it will appear that they have "returned to normal" in energy and behavior, but their physiology will have changed in an important way, altering their *needs* (e.g., for dietary iron) and *capacities* (e.g., ability to absorb atmospheric oxygen). This change is the foundation for altitude training for endurance athletes—by returning to low altitude competition from time at high altitudes, they have a greater capacity to provide energy to their muscles than those who remained at low altitude. Until, that is, they re-accommodate to low altitudes. Because accommodation is not the return of the body to an endogenously given set of parameters after a perturbation, but the alteration of the body to make normal activity possible in a new environment, this is quite a different process from sensory habituation.

Similarly, individuals continually living on the margin for income, wealth, or employment psycho-physically accommodate the need for continually elevated attention to small, everyday gains or losses for themselves or their family, and for continued motivation to find a way out, by maintaining high levels of cortisol in the blood. It would hardly be functional for such individuals to simply "get used to" their precarious positions by paying the same level of attention to marginal gains or losses as do most of us with regular employment, medical insurance, and average-or-better salaries. The accommodation that low-income individuals make by experiencing heightened stress thus has a functional basis in adapting their state and dispositions to their particular circumstances, even as it exacts a continuous toll on them mentally and physically.

'Accommodation' is typically used for middle- to long-term adaptation, but similar processes occur at much shorter temporal scales. When your eyes change shape to keep in focus an object moving toward you, this, too, is accommodation. Notice that this is not a mere reflex, like blinking. Reflexes are stereotyped, relatively invariant responses to immediately *past* stimuli. Keeping a moving object clearly in focus involves both accommodation and "smooth" eye movement. What drives the amount of tensioning by eye muscles in visual accommodation and smooth eye motion is computed information about where the object is expected *next*, given its current trajectory, along with information about how sharp the edges or details of this object are expected to be, how it has been changing in visual angle, etc. This is an information-intensive system, and one that infants have to learn to be good at—at birth, they are quite myopic.

From moment to moment, one is faced with the need to adapt to one's environment relative to one's needs (avoiding a moving object, or keeping one's family fed) or goals (tracking one's prey, or finding a job). To manage well one's circumstances and prospects, individuals must not only be equipped for persistent changes in levels of attention, effort, and satisfaction (as in chronically-elevated stress) but also more rapid reallocations of attention, effort, and satisfaction. Individuals who were simply cast down by adversity, who could not experience reward in a finding a temporary solution, pleasure in being with one's children or making a social contact, anger at a slight, amusement in a joke, sympathy for another's loss, or excitement about a new prospect, would be insensitive to many features we must be alive to in order to function well as complex, social beings. That is, to meet life's challenges or achieve human

aspirations, even in a marginal or precarious situation, once must retain capacities for wide emotional responsiveness and maintenance of self even as one struggles. A 2010 poll found:

Table 2. Positive and negative emotions experienced during the previous day, US, 2010 (Gallup-Healthways Well-Being Index, 2010)

	Worry	Sadness	Stress	Happiness	Enjoyment	Smile or laugh
Employed	28	12	40	91	87	85
Unemployed 1 month	40	25	42	89	88	83
Unemployed 1-6 months	46	27	52	87	85	81
Unemployed > 6 months	55	34	54	82	79	74

Individuals unemployed over 6 months, despite experiencing on average twice the level of worry and sadness of those with jobs, experienced on average much less dramatic reductions in happiness, enjoyment, or amusement, retaining an average level of experience of happiness and enjoyment at approximately 80%. When these more episodic forms of positive affect become chronically unavailable, the result is not a more adaptive accommodation to a demanding situation, but depression and severe problems coping.

Thus we need to distinguish, in affect, between more or less persistent changes in underlying state, such as chronic anxiety, on the one hand, and episodic sensitivity in affective tone or sense of satisfaction, on the other. Adaptive affective systems are capable of both, so that "subjective well-being" becomes an amalgam of different kinds of responsiveness to one's condition and its prospects. Each is supplying a different kind of information to the individual with different roles in shaping behavior—more persistent information about success or failure in meeting fundamental life concerns, calling for chronic vigilance vs. more episodic information about one's immediate circumstance, calling for a wide register of positive as well as negative responses.

These different roles help us to explain what otherwise is a puzzling feature of subjective well-being, namely, why the subjective well-being of those struggling to get by in relatively less prosperous societies is higher than the level of subjective well-being of those struggling to get by in more prosperous societies. Unemployment and marginality in a society in which employment and relative prosperity are the norm communicates different information to the individual about what he or she is actively failing to do, in comparison with unemployment or marginality in a society where this is the lot of the great majority. When unemployment and marginality are experienced as *personal* failure the level of stress is correspondingly high, as measured both by subjective reports (Table 2; Piachaud *et al.*, 2009) and blood cortisol (Voss, 2004; Maier *et al.*, 2005).

In less prosperous societies, in which poverty and unemployment or underemployment are pervasive, we see a somewhat different affective profile for those who are barely managing on meager resources. Consider:

Table 3. Percent satisfied with standard of living, thriving, struggling, or suffering: three nation comparison (Gallup World Survey, 2012; World Bank 2005-11)

	Satisfied with standard of living	Thriving	Struggling	Suffering	Per capita GDP
Bangladesh	74%	13%	76%	11%	1,777
Namibia	61%	11%	79%	10%	6,801
Thailand	83%	37%	60%	2%	8,646
United States	78%	58%	38%	4%	48,112

In Bangladesh three-quarters of the population place themselves on a ladder scale in the "struggling", and only 13% in the category "thriving". They have not, then, habituated to their poverty, or formed "adaptive preferences", or judged themselves to be thriving by means of social comparison. Yet three-quarters also report themselves satisfied with their standard of living. This suggests awareness of the distance between their existence and true thriving, yet this does not single them out in a counter-normative way as possessing unsatisfactory lives. If struggling is the norm, rather than a personal failure, and if they *succeed* in getting by despite the daily challenges, that is a form of satisfaction with one's is managing to achieve by way of a life. It would not be informative in such a setting for one's affective system to signal that one is *failing* in life for want of thriving, since there are not available paths to attain thriving. Yet in a more prosperous society, where the majority find paths to thriving and "just getting by" is not the norm, it *is* informative—if nonetheless brutal—to be kept aware of this, and the potential it represents to the unemployed individual.

At the same time, such satisfaction at struggling and succeeding should not translate into *complacency* or *insensitivity* to changes in precariousness. Haushofer *et al*. (2012), studying Kenya, found that weather effects that *increased* the precariousness of subsistence farmers increased both reported subjective stress and blood cortisol, while Fernald and Gunner (2009) found that children living in poverty in Mexico whose mothers were able to participate in a program that *lessened* precariousness showed lower cortisol levels than children from comparably poor families not participating in the program.

Even in very marginal conditions, adaptation requires sensitivity to levels of risk, whether moment to moment, or over time. At the neuro-physiological level, the mouse nucleus accumbens, which plays a critical role in reward and behavior regulation, appears to allocate greater or lesser portions of its "affective keyboard" to positive vs. negative stimuli as a recalibration to a given level of environmental risk (Reynolds & Berridge, 2008). This permits it to retain sensitivity the changes in risk or relative degrees of risk, whether overall risk is high or low, so that the individual is able to make an appropriate allocation of mental and physical resources.

A colleague, remembering portions of her childhood was spent in a war zone in Lebanon, explained to me that, despite the war, she still had to get to school, run errands, play with friends, find a boyfriend. As she put it, just as we consult the weather or the traffic when we go outside, so did members

of her family learn to for the sound firing or explosions to decide which streets to take, or which errands to rush to complete or postpone. She still had to do homework, worry about how she looked, and get along with her parents. Breaking up with a best friend or making a new friend still had to matter, even though she might learn the same day that a cousin had been killed in the cross-fire.

Revisiting national data

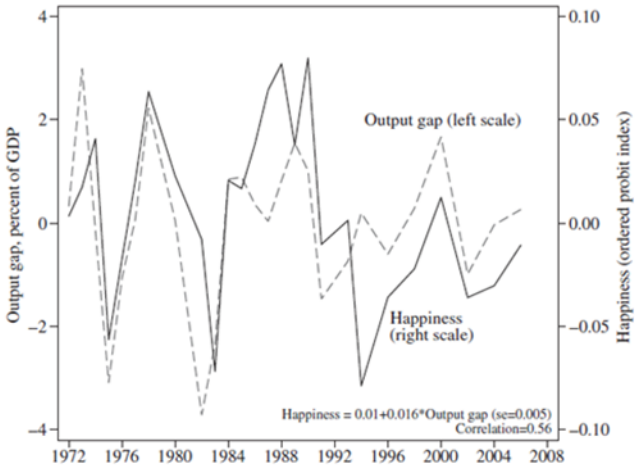
A lower GDP per capita thus might not be associated with lower "satisfaction" with the standard of living, as the comparison between the Thailand and the US suggests, and very large differences in GDP per capita need not translate into very large differences in satisfaction with one's standard of living, as the comparison between of Bangladesh and Namibia with the US suggests. These statistics, like the statistics for across a wide range of countries in Fig. 8, suggest that adaptation of one's expectations for a "satisfactory" standard of living to the general level of prosperity of one's surrounding society tends to occur, even though this is not habituation to one's circumstance. The juvenile squirrel is no longer "struggling" to contend with pedestrians, and you are no longer "struggling" with your sense of the conspicuousness of your solar panels to the neighbors, or yourself. But the majority of the people of Bangladesh, Namibia, Thailand know themselves to still be "struggling".

At the same time, Table 3 suggests that greater per capita GDP might not translate into a markedly higher sense of satisfaction with one's standard of living. Does this mean that well-being itself is not increased? By lowering *typical* precariousness, and thereby lowering the rates of struggling and failing that result from precariousness, lives will have been improved in dimensions that *do* matter to people—though in ways that have a complex rather than linear connection with expressed satisfaction with one's standard of living or life as a whole. (We will see a similar effect in studies of disability, below.)

Subjective well-being as information and guidance

The foregoing observations encourage us to think about questions of the relation of income or financial situation to subjective well-being as involving both longer-run and shorter-term adaptation or accommodation, via fundamental psychic processes that play a coordinated role in *informing* and *regulating* the organism's responses to its environment. Response of subjective well-being to economic changes may not be highly visible when averaged over long periods of time, but it can be visible in the small in terms of responses to variations in whether economic activity is increasing or decreasing (e.g., as measured by the gap between output and available capacity):

Fig. 14. Subjective well-being and the output gap, US 1972-2006 (General Social Survey)



These are relatively finely-calibrated changes, a matter of a change of .05 up or down. What they suggest, however, is a highly non-random fluctuation in subjective well-being, closely attuned to fine gradations in economic activity—in this case with a small phase shift because this indicator is most likely to affect individual lives materially rather than via an informational route (in contrast, perhaps, to other, more publicized, economic indicators).

This fits subjective well-being within an emerging conception of emotion generally as functioning to inform and regulate. As Phoebe Ellsworth, a cognitive social psychologist, and Randolf Nesse, an evolutionary psychologist and clinical psychiatrist, write:

> Emotions are modes of functioning, shaped by natural selection, that coordinate physiological, cognitive, motivational, behavioral, and subjective responses in patterns that *increase the ability to meet the adaptive challenges of situations* that have recurred over evolutionary time. [Nesse & Ellsworth, 2009]

It has become clear that affect is *continuously* operative, not just in arousal states such as fear or surprise (Izard, 2007). Thus, the insula appears to carry out a continual, real-time synthesis of information from primary interoceptive and homeostatic body systems, environmental conditions, and social, goal-seeking, normative, and affective systems, which functions as a sensitive indicator how well or ill the individual is doing overall (Craig, 2009). This evaluative information is in turn used to guide aspects of attention, cognition, "felt experience", motivation, and action.

Affect, then, should be seen as a continuous process that combines perception and evaluation, in a way that enters the stream of perceptual processing early (within 100 ms.) and encodes information in evaluative terms and degrees—as favorable or unfavorable, certain or uncertain, urgent or status-quo, relevant to self or others, indicating risk or reward, and so on (for a more extended discussion, see Duncan & Barrett, 2008; Seligman *et al.*, 2013 and Railton, 2014, 2018). This encoding then can prime and help shape a suite of other responses, including attention, memory, cognition, motivation, and action-readiness. The affective system is critical to such functions as spatial representation, mapping and simulating alternative courses of action, empathy and theory of mind, social evaluation, and motivation.

Although it is the product of natural selection, it appears to be flexible and adaptive, drawing widely upon information in the brain. Indeed, the affective system is the primary locus of experience-based learning and memory. All this has special relevance for thinking of motivation: rather than conceive the mind as a bundle of basic motivational drives, most motivation appears to be *downstream* from affective evaluation, so that the allocation of mental and bodily resources is directly shaped by current explicit or implicit assessments of the types of value and degrees of urgency or uncertainty present in a situation. In some circumstances individuals may respond by paying primary attention to the immediate needs of the self, in other situations, by paying primary attention to the needs of others, or long-term, abstract goals or values.

The principal elements of subjective well-being—including feelings of life-satisfaction and positive or negative mood, as well as a sense of striving or struggling, or of satisfaction and dissatisfaction in pursuing one's goals—belong in this family of affective attitudes that function to help attune and motivate human responses in a coordinated way. Dramatic changes in one's circumstance that call for departure from "business as usual" tend to produce *aroused* affective states such as fear, surprise, anxiety, anger, joy, or excitement, and reallocating mental and physical resources accordingly. But circumstances more congruent with one's expectations and making progress toward one's aims tend to produce *non-aroused* affective states such as confidence, calm, satisfaction, positive mood, trust, and assurance. This array of responses enables humans to respond aptly to circumstances, thus measurements of subjective well-being should be expected to be informative as to the condition of the individual and how his or her circumstances, actual or projected, relate to her needs and values.

Not all affect is aroused, however. For example, in humans a default level of confidence, trust, and interest function to sustain an individual's active engagement with the world and other people, promoting active learning through taking in information to effortlessly update beliefs on the basis of experience and testimony. The immediate mental and physical effects of *positive* emotional states like confidence, trust, liking, interest, and excitement are *acceptance* (e.g., openness to new experiences, new people, new courses of action) and *approach* (e.g., exploration, relationship formation and maintenance, willingness to share information and act cooperatively). The immediate mental and physical effects of *negative* emotional states like fear, distrust, sadness, disgust, and hatred are *rejection* (e.g., closure to new experiences, new people, new courses of action) and *withdrawal* (e.g., defensiveness, relationship disinvestment or rupturing, unwillingness to share information or act cooperatively). A wholly neutral emotional state would dispose neither to accept nor reject, neither approach nor withdraw.

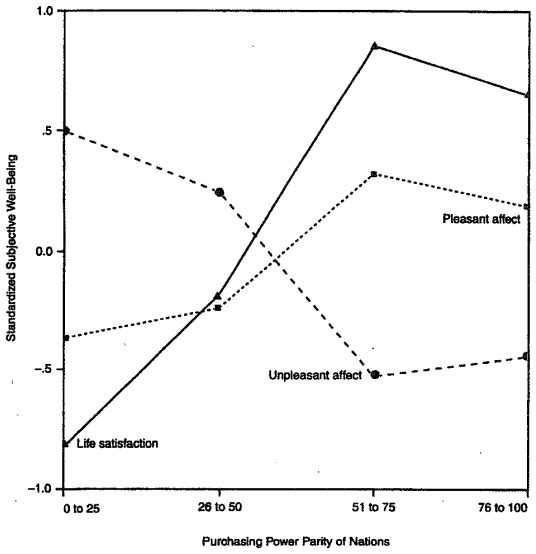
This suggests an explanation, lying in the fundamental dynamics of affective processes, for the high prevalence of positive subjective well-being in most people most of the time. Were the characteristic human default state overall negative, individuals' predominant stance toward the world would be rejection and avoidance—a pattern we see in individuals suffering chronically high levels of negative affect. Were the characteristic default state overall neutral, individuals' predominant stance would be indifference and non-engagement. Only if the default state is on the whole positive will individuals' predominant stance be one of approach and acceptance. Approach and acceptance are not, however, ways of *insulating* oneself against negative information. On the contrary, they open the individual to *gain* and *take seriously* the information available to them in their environment—we are foragers, and descended from foragers, and this includes information. Approach and acceptance fuel forward projection and feedback by discrepancy-detection, which is increasingly seen as the foundation

for learning in intelligent animals (Schultz, 2008). The key is that *default* approach and acceptance make such learning possible, while default withdrawal and rejection, or default inactions and indifference, do not. A "set-point" in the positive range, then, is not an example of human lack of realism, but an integral part of how our affective system supplies the motivation and coordination of faculties that promotes effective learning, of bad news as well as good.

But now notice, if individuals *could not* in the main accommodate to a reasonable degree their expectations and sense of well-being to their horizon of actual possibilities, then life would be a continual accumulation of negative experiences, disappointed expectations, frustrations, losses, and failures. They could not sustain the levels of positive affect needed to be effective at gaining and using information projectively to guide their behavior toward the goods, and away from the harms, that *are* available to them through intelligent effort. At the same time, they would not be effective at gaining and using information in these ways if their system were blunted in its sensitivity and responsiveness to everyday gains and losses—if, in being "satisfied", they lose the ability to "struggle" effectively.

We should therefore expect that, even if individuals' overall life satisfaction does not in many circumstances prove an exciting social indicator to watch, there should remain considerable variability in the experience of positive and negative affect over the course of daily life, reflecting subtle or large changes in prospects that provide feedback about whether we are doing better or worse in deploying our resources to contend with our circumstances to realize our aims. Recall Figs. 13 and 14. And consider in this light Fig. 16:

Fig. 16. Positive and negative affect and life satisfaction vs. national average purchasing power



Increased average income or purchasing power can very significantly reduce average negative affect and increase average positive affect as income rises from a very low level, but once a level of reasonable sufficiency is attained, the power of income to reduce the sources of negative affect in life, or provide a source of positive affect, is much attenuated.

Why would that be? Think of the actual sources of positive and negative affect over the course of one's typical day or a week. If we compare a household with annual income of $90,000 to one with an annual income of $1,000,000 or more, do those with much higher incomes have fewer problems with their children or spouses? less conflict or pressure at work? better friendships? less status anxiety? fewer doubts about whether a given decision or purchase or remark was the right one? fewer frustrations in trying to organize time and effort in order to accomplish in the course of the day or week or month what one hoped to accomplish? It is not hard to believe that those with higher incomes simply experience higher-income versions of essentially the same sources of pleasure and annoyance, satisfaction and anxiety or frustration, as those with sufficient income. They must, like all of us, read these signs of making, or failing to make, headway toward goals, or progress in problem-solving, or good decisions about time allocation or consumption, or appropriate responses to a family crisis or a child's difficulties in school, or the right gestures toward others, …, even if their income or overall level of life satisfaction is high. For example, Norbert Schwarz has suggested that higher income often comes with more demands upon one's time and more commitment to work-related activities, and this tends to offset any gains in the quality of non-work activities (Schwartz, personal communication; see also Quoidbach, *et al.* 2010). Visitors to the US are often surprised to see how much Americans forgo or compromise in the rest of their lives for the sake of higher income.

If subjective well-being is indeed about accommodation and adaptation, via the capacity of affect to inform and regulate, what might we expect?

Where "habituation" and "return to a set point" do not occur, or occur slowly and incompletely

Let's briefly survey some evidence concerning life conditions that appear to register in a fairly persistent way in one's overall satisfaction or positive and negative affect. Recall our discussion at the outset of aging and health, where we saw that an accurate sense of one's deteriorating health condition need not lead to a negative subjective well-being in countries with good social services for the elderly. This makes it clear that people do not simply *ignore* their objective condition, but subjective well-being can be sustained if one is doing reasonably well *given* one's health condition. As people's horizon of prospective possibilities changes with age, the positive or negative information embodied in health changes, with associated effects on one's sense of well-being. For example:

Table 4. Mean happiness by self-reported health status and birth cohort, US, 1972-2000 (General Social Survey, Opinion Research Center, 2002)

	Mean happiness			
Birth cohort	Excellent health	Good health	Fair health	Poor health
1951–1960	2.36	2.12	1.85	1.63
1941–1950	2.37	2.17	1.92	1.74
1931–1940	2.43	2.23	1.98	1.74
1921–1930	2.48	2.24	2.06	1.83
1911–1920	2.52	2.27	2.12	1.96

As one ages, excellent health becomes an increasingly strong sign that one is doing well in attaining one's aims, and poor health an increasingly weak sign that one is doing something wrong. This should lead to a general elevation of subjective well-being in response to the various steps in the progression from poor to excellent health, and that is indeed what one sees in Table 4. Such *differences* in subjective response to a given level of health require that individuals be explicitly or implicitly *aware* of their personal level of health. If, with habituation, they came to *ignore* it, they could not remain alive to the countless and indefinitely-varied adjustments in their behavior or planning required by their current level of mobility, muscle strength, endurance, speed of reaction, and ability to keep or restore balance. They could not effectively *adapt* to their changing circumstances.

An interesting implication is that when a health-related cost is perceived as temporary and changeable, rather than long-term and irrevocable, individuals should be *less* satisfied with their condition—like the unemployed, they need to be motivated not to remain in their condition. This is what Smith *et al.* (2009) found by comparing patients with temporary as opposed to permanent colostomy sacs, and finding that those with temporary sacs expressed greater dissatisfaction with their condition. Or consider:

Fig. 17. Change in average subjective well-being: Russia, 1981-1995

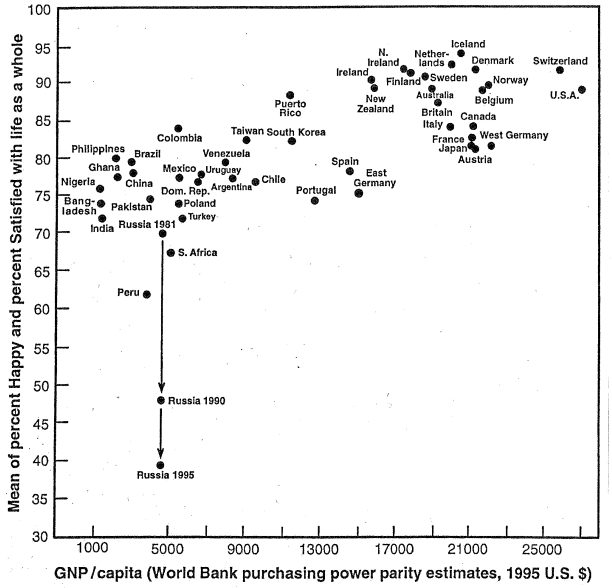


Figure 7.4
Collapse of communism and the decline of subjective well-being in Russia. The correlation between GDP/capita and subjective well-being, omitting the former communist societies, is $r = 0.74$. Source: World Values Surveys. Data for Russia in 1981 from Tambov oblast, 1981.

(As conditions have stabilized, subjective well-being has recovered somewhat in Russia since 1995, though it has not reached prior levels, OECD, 2017.) It is one thing, evidently, to "struggle" with insecurity and precariousness in Bangladesh, where this has been the lot of the great majority of the population for as long as anyone can remember, and when most adults "learn the ropes". It is another thing to struggle with a new insecurity and precariousness in a country the rules of which have fundamentally changed, and where many of the normal supports of life have suddenly been withdrawn, so that individuals cannot take an underlying stability in their situation for granted.

Let us look at a range of conditions to get an idea of which sorts of losses, costs, or conditions seem to have a persistent effect on subjective well-being.

Table 5. Demography and subjective well-being: Citizens and foreigners living in Switzerland (Leu, Burri, and Priester, 1997) – 10 point scale

- Higher education 8.41
- Married 8.36
- Widowed 8.35 (with partner)
- Self-employed 8.31
- Swiss citizen 8.30
- Retired 8.23
- Employed 8.21
- Male 8.22
- Female 8.22
- Single 8.17 (with partner)

- Widowed 8.16 (without partner)
- Student 8.16
- Single 8.01 (without partner)
- Low education 7.97
- Divorced 7.90 (with partner)
- Foreigner 7.62
- Active bad health 7.48
- Divorced 7.43 (without partner)
- Separated 6.62
- Unemployed 6.56
- Separated 6.33 (with partner)

Table 5 makes it clear that certain conditions have a much more lasting and serious effects on subjective well-being than others.

This draws attention to the regulative as well as informative aspect of affect and a sense of life satisfaction. The negative effect of being widowed or divorced, for example, is much less if one has been able to respond to one's condition by finding a new partner. Put in regulative terms, as long as one lacks such a partner, one is motivated to find one by a lower overall sense of satisfaction or well-being; when one has found a new partner, this lack for has been met. Similarly for single individuals who have, vs. have not, a partner. Sharing the bottom rung of the ratings with the unemployed, those separated are in an unresolved, marginal position that they must be motivated to change.

Being in a stigmatized group within the larger population—in Switzerland, for example, being a foreigner or not well-educated—leads to a daily experience of failing to some degree to meet social expectations, and thus one does not simply "habituate" to this condition. Members of marginalized and stigmatized groups must make countless everyday adaptations to the ways they are perceived by others, and the differences in the prospects they face in new situations. Marginalized and stigmatized groups can thus be found to have lower levels of subjective well-being across the globe, and, like the unemployed (and especially the long-term unemployed who continue to seek work, see Elliott & Dockery, n.d.), to show higher levels of stress with poorer associated health outcomes (Cohen *et al.*, 2007).

Normative expectations matter, and, for better or worse, are highly relevant to how an individual is to comport herself in the world. Frey and Stutzer (2002) consider data from a number of highly developed countries and find that unemployed workers suffer exclusion, depression, anxiety and social stigma. And Lucas *et al.* (2004) find in a 15-year longitudinal study of Germany that individuals who have lived through unemployment do not return to their pre-unemployment "set point" once rehired, nor do they (like animals habituated to a stimulus) react less strongly to unemployment if it occurs again.

We can compare unemployment with widowhood. Recent death of a spouse can be devastating, but being widowed, whether with or without a new partner, is a naturally-occurring fact of life and not a socially stigmatizing situation. To non-stigmatized set-backs, it appears, adjustment can take a different course. While it might take as long as 3-8 years, widowed individuals do return to typical levels of

subjective well-being—but stigmatized individuals or groups can remain in a society indefinitely without losing their deficit in subjective well-being.

Table 5 gives us another counter-normative contrast: divorce, unlike widowhood, tends to cast an enduring shadow on one's subjective well-being, even when one eventually finds a new partner. According to *Roper Reports* (1979, 1995), the percentage of the population wanting a happy marriage is *higher* among those who are never married (65%), divorced (63%), or widowed (62%), than in those who are married, but unhappily (56%) (Easterlin, 2003; Carr, 2004). Falling short of this ideal is closely related to the fundamental need for social recognition and legitimacy.

And the need for affiliation. Indeed, evidence suggests that isolation—not in the sense of living alone, or interacting infrequently with others, but in the rather precise sense of lacking a companion or "confidant"—is a more potent predictor of subjective well-being and physical and mental health than being married, or having children, or being widowed. Affiliative needs are real, and inability to meet them does not result, it seems, in "habituation" to being cut off, but in continuing lowered satisfaction with one's life (Chappell & Badger, 1989). Consider:

Table 6. Life evaluation, positive and negative affect, and stress: ratio of regression coefficient to *log*(income) regression coefficient, US 2008-2009 (Kahneman & Deaton, 2010)

	Positive affect	Blue affect	Stress	Ladder
Children	0.08	-0.37	-2.47	-0.11
Care-giver	-0.49	-1.02	-2.99	-0.25
Health cond.	-1.36	-1.22	-3.15	-0.48
Headache	-4.45	-3.41	-9.82	-0.78
Alone	-7.13	-2.10	-3.73	-0.75

'Care giver' = daily care for elderly or disabled family member; 'Health condition' = chronic condition such as heart or circulatory disease, asthma, or cancer; 'Headache' = currently suffering headache; 'Alone' = no social contact of any kind, including telephone or e-mail, during last 24 hours. A value greater than positive or negative one means this factor has a greater effect on the relevant category (positive or negative affect, stress, life satisfaction) than being in the portion of the population earning more than $48,000 annually.

The ratios in Table 6 make it clear how much isolation influences both more chronic accommodation, such as stress and life-satisfaction, and the quality of daily experience, as reflected in the strong negative co-efficient with positive affect (blue affect is "reverse-coded", so there is a significant effect in increasing the "blueness" of experience as well). And, like Tables 1 and 2, Table 6 shows the *interconnection* of the various elements of the affective system—short-term adaptation and longer-term accommodation are two aspects of a shared capacity to "attune" oneself to one's condition and its prospects and problems. This is why local effects (such as a headache or situational cues) can so effectively alter one's sense of *both* dimensions of subjective well-being.

There are many dimensions to how well our lives are going, but two are especially salient. (1) How well, from moment to moment, we are doing in meeting our current needs, achieving our near-term ends, connecting with those around us, solving a pressing problem, or getting ahead in an on-going competition—to this corresponds, predominantly, the *affective* dimension of subjective well-being. (2) How well, from a more global perspective, we are doing in achieving such life goals as material security, creating and providing for a family, achieving the respect or love or companionship of others, making meaningful contributions or accomplishments, realizing long-term ambitions such as the acquisition of wealth, power, or renown, and so on—to this corresponds, predominantly, the *overall life satisfaction* dimension of subjective well-being. The first has to be a fast-responding, sensitive indicator that helps us steer our efforts toward success and away from failure or frustration; the second has to be a running idea of how well we are succeeding in a more global sense, regardless of local variations. Both are needed for effective sensitivity to the world and regulation of one's responses to it.

This is sound design if affect and the experience of subjective well-being is at heart a system for informing and guiding individuals—for *emotional intelligence and responsiveness*—rather than for handing out prizes for absolute life success. As a form of living intelligence, we would not expect that

this system attaches much significance to one's level of material well-being once this has exceeded a level such that material gains have done all they can to help us with the challenges we typically face in life: to meet basic needs, to promote health, learning, and security, and to find affection, affiliation, social support, respect, autonomy, and meaning. In realizing these aims, money can accomplish only so much.

What we should expect instead is that our sense of subjective well-being should help keep us alert to continuing challenges and unmet needs, especially when these call for daily focus and adjustment in thought, feeling, and action (such as precariousness, social stigmatization, relationship failure, and unemployment). Moment to moment, it should help us to discriminate different kinds of challenges and opportunities, and help support functional responses to them.

The "readings" obtained by measuring subjective well-being thus can be expected to present a coherent and informative picture of what individuals are living through and how they are managing with it. This information is vital for personal and social decision-making, and, more generally, for understanding what people's lives are like to live. Precariousness, and associated lack of material or social resources to meet life's challenges, emerge as especially important. Among those countries that stand out in subjective well-being internationally Figs. 1 and 8, a notable feature is not genetic similarity or high material wealth but similarity with respect to the reduction of precariousness, either by social resources such as extended kinship and community relations (as in Latin America), or by social resources in the form of institutional guarantees (as in Scandinavia)—above a level of reasonable sufficiency, absolute material level seems much less important. Utilitarians and others concerned to improve objective well-being might use the information contained in subjective well-being surveys to identify the most effective ways to make a difference. The result might dissociate Utilitarianism from the idea that the whole world must be "brought to the table" of high consumption levels present in the contemporary US. Rather we might all live on more modest material budgets, it seems, and thus do the planet and future generations a favor—without sacrificing the *subjective* dimensions of life satisfaction and affective balance. If, that is, we create the conditions in which individuals have adequate support to meet their needs, escape precariousness, achieve social connection, and find opportunities for self-development. Political freedom and civil liberties, too, matter. Though subjective well-being is not a linear function of income, it is nearly a linear function of a measure of political freedoms and civil liberties:

Fig. 18. Subjective well-being vs. civil liberties and political rights: 1990-1995.

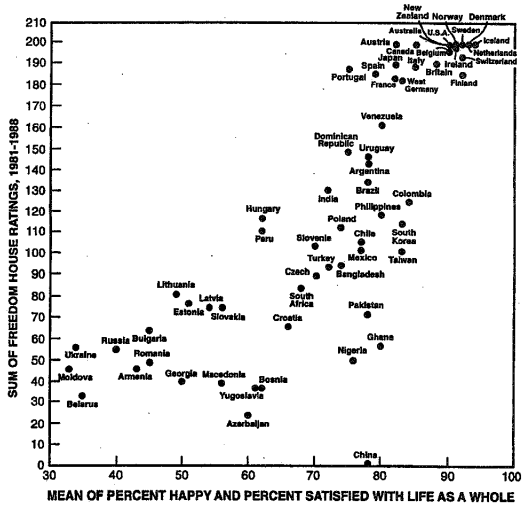


Figure 7.5
Subjective well-being and democratic institutions ($r = 0.78$, $N = 62$, $p = 0.0000$). The vertical axis shows the sum of the Freedom House ratings for civil liberties and political rights. Since these ratings give high scores for low levels of democracy, we reversed polarity by subtracting these sums from 236 (China, which had the maximum score of 235, has a score of 1 after this transformation). The horizontal axis reflects each public's mean factor score on happiness and overall life satisfaction and subjective well-being. Source: Freedom House surveys reported in successive editions of *Freedom in the World*; survey data from the 1990 and 1995 World Values Surveys.

Securing such liberties and rights is an important social accomplishment in itself, and an indicator of much else about how lives are going within the society. The most developed forms of civil liberties and political rights are typically found in relatively prosperous countries, though it is clear from the shape of the main trend in Fig. 18 vs. Fig. 8 that this effect is not entirely attributable to relative affluence. Moreover, cause and effect are likely to be highly confounded here. Importantly, as the presence of a number of relatively less affluent countries high on the scale of civil liberties and political rights, creating the conditions in which people can widely enjoy civil liberties and political rights does not in itself require that we impoverish future generations by chewing through the world's scarce resources to support high levels of consumer culture. Interestingly, Utilitarian claims that overall utility can be raised by increasing civil liberties and political rights, and lowered by taking them away, would appear to receive some confirmation, after all.

At the same time, the evidence from the data on subjective well-being undermines a certain conception of "adaptive preferences" or "hedonic adaptation". That people in Bangladesh adjust their expectations to their horizon of possibility, and thus do not find high levels of satisfaction in a life of successful struggle to make ends meet, raise a family, and maintain one's standing in a community despite poverty, does not suggest that they confuse their condition with the best sort of life, or genuinely thriving—that they would not prefer less precariousness and more resources and opportunities. Similarly, those with permanent colostomy sacs might show less dissatisfaction than those with temporary sacs, and all patients might show less dissatisfaction with the arrangement than they had imagined beforehand. But all also retain a preference, if wishes were so, to live without need for a sack.

Moreover, people in the wealthiest countries do not appear to have acquired "expensive tastes" the satisfaction of which makes them ever-so-much-more happy or satisfied than those in less wealthy

countries whose less expensive tastes have been satisfied. Only one measure of one dimension of subjective well-being—ladder life satisfaction—suggests that very large incomes raise any aspect of subjective well-being over large incomes. And even then the gain is so small as to require a doubling of income to find any noticeable effect, and appears to decline marginally as incomes become very, very large.

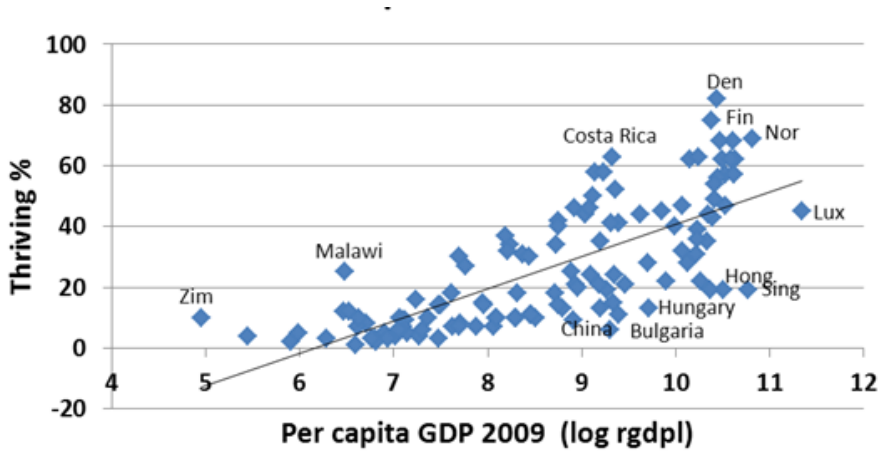
Subjective well-being is not a straightforward indicator of objective well-being, but not because it habituates or functions primarily to boost one's self-image. Rather, it is not a direct indicator because it is an intelligent system that helps us accommodate adaptively to widely varying conditions. If typical subjective well-being is high this in part reflects the orientation needed for evidence-seeking trial-and-error learning, and for effective motivation and behavior in response to that evidence. This is not a measure of the objective *level* of one's existence, then, though *variations* in subjective well-being may be sensitive indicators of important aspects of one's life conditions or success in meeting them. And it also is why we should not be surprised if, even in quite poor countries, if people are able to cope they experience reasonably high levels of subjective well-being. Very low subjective well-being amounts to depression and a dysfunctional insensitivity to one's possibilities and prospects, diminishing expectations and motivation indiscriminately—rather than a realistic appraisal of, or adaptation to, marginal life circumstances.

And, finally, what of the enduringly high reported subjective well-being in the Scandinavian countries? Consider Denmark. In the US, with a GDP per capita 20% higher than Denmark, 58% of the population categorize themselves as thriving, 38% struggling, and 4% suffering. In Denmark, 74% report that they are thriving, 24% struggling, and 2% suffering (Gallup, 2011). It was not added wealth that enabled so many more Danes to thrive, and to escape struggling or suffering. Comparing the two societies, one might notice the very salient fact that Denmark seems to have made possible for a wider swath of its population the meeting of basic needs, security from precariousness, and the social preconditions for enjoyment of such goods as connectedness, mutual respect, and self-development. If so, then there is no magic in this formula—to the extent that we see elements of it in other societies around the globe, we see higher levels of reported subjective well-being. And it is a formula that accords well with our notions of what makes a life go well objectively. People's responses to the much-derided "happiness surveys" of subjective well-being might be telling us something important about their objective well-being after all—if we will but listen carefully enough to what they are saying.

Subjective well-being and climate

One suggestion to draw from these considerations is that attaining high levels of subjective well-being within a population need not be as resource-intensive as one might have imagined. Consider:

Fig. 19. Relationship of thriving and per capita GDP



And also:

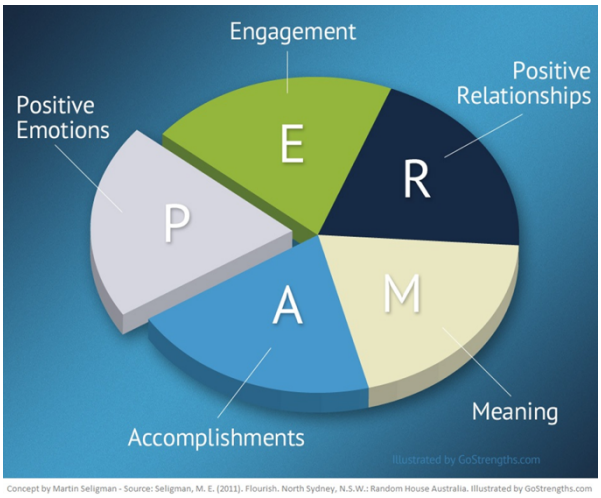
Fig. 20: Carbon footprint per capita (in metric tons, World Bank 2014)

	1980	2013
United States	20.8	16.4
Denmark	11.8	6.8
Sweden	8.6	4.6
Costa Rica	1.0	1.6

Clearly there are far too many variables in play to compare the US with Denmark, Sweden, or Costa Rica directly—for no other reason than questions of size and diversity of population. But the point is not direct comparison, rather, to examine closely the factors in virtue of which, for example, Denmark can reliably achieve much higher rates of thriving, and lower rates of struggling or failing, than the US, despite a persistently lower GDP per capita. Or to understand the adjustments made in the economy and daily life in Denmark or Sweden to lower their carbon footprints by almost half since 1980. Or to appreciate the contribution of the kinds of *social resources* in virtue of which both Denmark and Costa Rica manage to provide a high level of average subjective well-being while consuming significantly fewer material resources per capita than the US.

Positive psychology is the study of the nature and sources of human well-being, and one well-established model is Martin Seligman's PERMA model:

Fig. 2. Elements of subjective well-being (Seligman, 2011)



The elements of PERMA (experience of positive emotions; engagement with absorbing activities that deploy a range of one's capacities and permit the development of mastery and "flow" experiences; relationships with family, friends, co-workers, community members, etc. that play a significant role in one's life; finding meaning in what one does or seeks, often with reference to a wider scheme of values that can underwrite a sense of contribution and self-worth; and accomplishment in the successful pursuit of one's goals or values) are not intrinsically *expensive, resource-intensive, restricted to a small portion of the population,* or *beyond the capacity of social support structures, education, employment policy, etc. to affect.* While material goods can contribute to the realization of these elements of a life, the level of material goods required in any given case need not be high. Of course, to the extent that individuals have materialist or positional values and goals (acquisition centrality, possession-defined success, the idea of acquisition as the pursuit of happiness, and a focus on relative economic standing as a measure of life-success), then material success or social position will play a significant role in what they take to have meaning or constitute accomplishment or important relationships with others. However, a fairly well-established body of research indicates that having materialist values, even controlling for a wide range of other factors, leads to lower subjective well-being *even for the person concerned* (Kasser, 2003) and that individual material consumption goods are less effective in raising subjective well-being than experiential or shared goods (Gilovich & Kumar, 2015; Caprariello & Reis, 2013). Indeed, even in purchasing behavior, which appears unsurprisingly to be more common at a given level of income in individuals with higher levels of materialism, seems to yield a less positive experience overall for the individual than for individuals with lower levels of materialism (Brown *et al.*, 2016).

Inhabiting a culture that encourages individuals to prioritize material consumption or social position, and to give these a central role in individual and social esteem and motivation at the expense of experiential, shared, social goods, or spiritual goods would seem to be a *risk factor* for individual subjective well-being. Materialist and positional values moreover have a feature of placing inherently high demands upon resources, and encouraging induced inequality—to be successful in these terms is to have significantly *more* than others or to have authority *over* others, so that these values cannot be jointly realized by a large portion of the population. Sustaining a system of materialist and positional values thus is also a *risk factor* for social levels of subjective well-being—there will typically be many "losers" for a smaller number of "winners". Focusing away from material and hierarchical values, individually or

socially, and shifting focus to values related to personal and familial development, social connectedness, and contributing to outcomes of shared meaning need not create more "losers" than "winners"—on the contrary, it can link one's own "winning" with others' "winning" as well, in ways that can strengthen relations, increase the meaningfulness of one's activities, and augment rather than undermine individual subjective well-being.

We've seen one potential source of materialist or positional values, in the "front-end effect" of gains in income and position discussed above. Another important source no doubt is cultural, and societies differ noticeably in this regard (Grouzet *et al.*, 2005). For example, affluent countries differ in terms of penetration of commercial interests into early childhood and schooling (Kasser *et al.*, 2007). But lack of resources and precariousness of one's material condition also conduce strongly to the prioritization of materialist values, so the extent of poverty in a country can also affect the pervasiveness of individual materialism. This of course makes sense, and we have seen that, when existence is precarious, subjective well-being and more objective health indicators suffer. This also means, however, that countries adopting social policies in which the lives of those less well-off are less precarious and unsupported—which need not be wealthy countries—constitute a social environment that is to that extent less encouraging toward materialist values. When this social environment includes support for self-development and education, for family life, and for choice in employment, the result can be to create a climate favorable to a wider array of intrinsic values.

So we have another "environment problem" to solve—the social environment that so deeply shapes the values and ideologies that, in turn, affect behavior. Coming to terms with this problem involves supporting social institutions and practices that can reduce precariousness and create a more favorable environment for the flourishing of values other than material concerns, and help individuals to acquire the wherewithal to realize more intrinsic value in their lives, individually and relationally.[124] Since such values are connected with relationships, meaning, and accomplishment, goals such as responding effectively to the threats of climate change and injustice could be candidates. Creating a world in which more individuals are able to achieve stable, relatively high levels of subjective well-being thus need not be antithetical to making changes required to contend with climate change or injustice—either in the nearer term or in the longer term. Of course, the interests arrayed on behalf of fostering consumption-oriented behavior in a society where so much of economic growth is driven by consumption are very strong, as are related ideologies. At some point, "the American Dream" came to be detached from a realistic picture of "the pursuit of happiness" and appended to a notion that happiness comes from abundant possessions and high income and social position—a dream, fueled by "rags to riches" biographies that, by their nature, not all can realize. But once we understand the role of affect and evaluation in guiding motivation, we can see that change in the "dream" is not an inevitable progression as egoism trumps social connectedness and larger-scale values. As values evolve, through whatever

[124] For example, an experiment in guaranteeing minimum incomes in the US, while controversial to interpret (Hum & Simpson, 1993), was seen by some interpreters as indicating that individuals, who on average were still eager to work, were able to search longer for, and find, work that better suited their skills and interests. A recent Finnish experiment showed, as a preliminary result: increased trust in others and in the political and legal system; gains in confidence about the future as well as subjective well-being and the reduction of stress; and less tendency to lose interest in various things previously considered enjoyable (Kangas *et al.*, 2019).

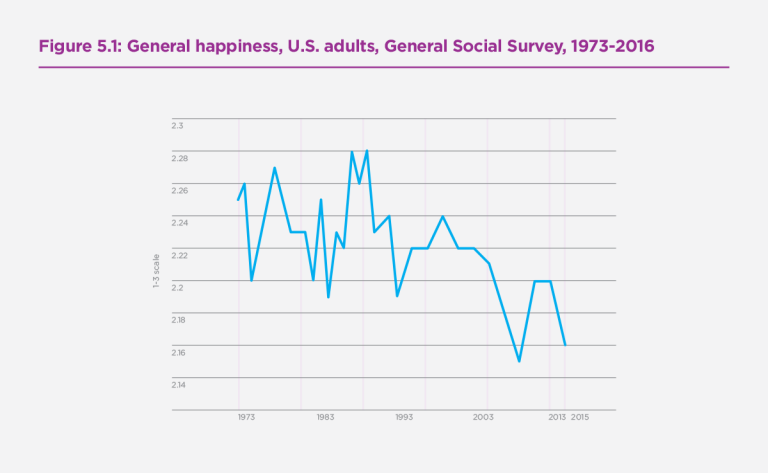
means, so does motivation, and the relative force of egoism is a variable, not a constant, in personal or social life.

Affect and valuation are central to learning, and learning can be on the side of recognizing that material and positional values are insufficient. As usually conceived, the "front-end effect" is a direct form of associative learning, connecting a raise in salary or gain in position with the experience of reward. But *why* would this be rewarding? And why would the reward be experienced simply upon *learning* of the promotion, not awaiting the arrival of the material changes it could effect in one's life? Humans are not simply dependent upon associative learning tied to proximate outcomes, and are notable for the extent of their capacity for *model-based learning*, enabling them to make connections between distant prospective intrinsic rewards and current actions. This is a foundation for human ambition, and the "front-end effect" may well owe something to it, since, for those attaching high *value* to material success, the news of a raise will be news of making progress toward that more distant value, and this will encourage more behavior of the kind that led to the raise. At the same time, however, this mechanism does not determine the *content* of more distant goals, since that will depend upon one's values. In a culture placing heavy emphasis upon material acquisition and social position, aims and ideals of these kinds will be highly salient and tend to attract admiration and effort. But if they do not live up to expectations, and fail to provide the valued social relations, or sense of meaningful accomplishment or fulfillment that they promise, these values can lose some of their luster. Over the course of a lifetime, most learn that pursuit of other values—family, friendship, social or spiritual involvement, the development of skills—can also promote subjective well-being, and perhaps do so more deeply or enduringly, and for more of us than the very few.

The values requisite for more forceful and politically effective concern with climate change and justice offer these features of caring for family responsibility, forming human connections in shared, meaningful enterprises, and connecting with larger purposes. As values sensitive to environmental concerns have become in fact more widespread, so has their capacity to motivate genuine engagement and effort—as we already see in many personal choices and social and political movement.

Moreover, at the same time, other kinds of learning are taking place. The technology for alternate energy and transport is becoming more cost effective than the traditional technologies they can replace, both in affluent countries and in the underdeveloped world (McKibben, 2019). And the effects of climate change become each year more severe and difficult to avoid in daily life—whether one is a struggling farmer facing drought, a fisherman seeing the displacement of species by rising water temperature, or the wealthy owner of a treasured beachfront house seeing uncontrollable erosion and increasing inability to find insurance.

The strong and reliable interest of people in promoting their subjective well-being does not in itself pose an obstacle to the social changes necessary to reduce the harms of climate change and injustice—on the contrary, these changes may be a key part of *enhancing* subjective well-being, now and in the future. Our current situation is *undesirable* as well as *unsustainable*—preserving our current "way of life" is not a way of preserving our happiness:



And making our way of life more sustainable might at the same time be making it more desirable. Strikingly, the developed countries that have taken the environmental challenge most seriously, and made the most progress in contending with it, stand above us in subjective well-being. It would not be an overall sacrifice to our subjective well-being to follow that path, though some large forces would like very much for us to think it would. Contending with the environment will require the imagination to think in terms of the well-being of future generations—but it may also require the imagination to shake off unsubstantiated but ideologically powerful ideas about the nature and source of our own well-being.[125]

[125] Acknowledgements. Fuller references to be supplied.

On what is "out of the question"

Lilian O'Brien

> **Drowning**
> Your spouse is drowning before your eyes. Another person who is unknown to you is also drowning. You can only save one at a time. You save your spouse. Once they are safe, you turn to save the other, but it is too late. Feeling shocked and guilty, you nevertheless think **(i)** "*I had to* save my spouse first." And in fact, in the heat of the moment you **(ii)** *didn't even stop to consider* the question of whom you should try to save first.

Such cases have been taken to highlight a tension between an agent's commitments and the allegedly impersonal demands of morality. (Williams, 1981) But these cases also highlight a tension between *commitment* and *self-government*. If we should be troubled by the question of how (or if) an intention formed much earlier could have rational or normative authority for an agent if she is self-governing (e.g. Ferrero 2010), surely we should be all the more troubled by the constraining hand of commitments of love. After all, such commitments exert much more "rational pressure", not to mind emotional pressure, than the average intention.[126] And when such commitments involve the **(i)** volitional incapacity and **(ii)** deliberative silencing that we see in Drowning, this tension demands scrutiny. (Frankfurt 1988; Watson 2004)

There are, I think, powerful considerations that speak against the view that commitments – and the commitments of love in particular - have an unassailable value. Here I argue for a limited claim. Although loss of normative control is more often associated with addiction, commitments of love have the strong potential to disrupt our capacity for normative control. More specifically, they are problematic because they undermine our capacity for practical deliberation and the epistemic and practical benefits that stem from this.[127]

1. The psychology of commitment

If we are to explore the tension between commitment and normative control, we must first gain some understanding of the psychology of commitment. This is the focus of this section (Section 1, (a) – (d)). As an under-researched species of "normative guidance" commitment is of interest in in its own right. (Railton, 2006, 2009)[128] The conflict with normative control will be returned to in Section 2.

(a) Action-based commitments and Role-based commitments

I will focus on cases of a person's commitment to her spouse, or a parent's commitment to her child. Such commitments are common, they are typically long-term, open-ended, wide-ranging, and complex. They differ in these respects from typical "*action-based*" commitments - commitments to perform some isolated action or series of actions, such as the action(s) of picking a friend up at the train station, or of

[126] This notion of "rational pressure" comes from Bratman 1989.

[127] The capacity for normative control will be understood as "… the capacity to govern oneself in accordance with one's values and the (usually diachronic) reasons one generates. While the capacity for intentional control over one's actions is required in order for one to count as an agent at all, normative control constitutes us as autonomous, self-governed individuals." Kennett 2013.

[128] There are interesting treatments of commitment in Chang, Shpall, and others. My aim is not to develop a theory of commitment, and so, I set any attempt to review such views aside. My main focus is on the impact that commitment has on a rational practical agent's capacity to exercise normative control. Consequently, I will focus on relevant features of the psychology of the committed agent. The kind of agent considered is a statistically normal adult human being.

getting to work on time every morning. Action commitments are usually not as open-ended and wide-ranging as role-commitments as they specify times, places, act-types, which give the agent a clear end-point for when they have complied with their commitment.

"*Role-based*" commitments, like those of spouse and parent, are typically held for a significant period of time, they involve clusters of obligations and entitlements, and the occupant of the role may accept that what is required of them may evolve in an open-ended way where she does not exercise much control over the exact nature and extent of what will be required of her.[129]

(b) Roles and practical standards

I will assume that the roles of, say, parent, spouse, teacher, friend, neighbour, project manager, are all *practical* roles. They can be occupied by agents, such as statistically normal adult human beings. The roles are defined or constituted by clusters of practical standards. Practical standards are standards of action – standards of thought and movement that agents must meet if they are to token certain act-types. To be the occupant of a practical role one must sometimes try to meet the practical standards that define or constitute the role, and perhaps also, one must sometimes non-accidentally meet these standards. For example, the role of parent has as a general constitutive practical standard: care for a child so that she remains healthy. This highly general constitutive standard gives rise to others, such as show the child love, feed the child, teach the child right from wrong, and so on. These practical standards will demand a role-occupants' compliance with yet other practical standards in specific contexts – cook dinner at 5pm, discuss some matter with the child after school, and so on. It is plausible to suppose, although it goes beyond my scope here to establish, that practical roles involve constitutive standards – standards that an agent must hold herself to if she is to count as occupying that role – and standards that are not constitutive, but which, when an agent holds herself to them, involve her acting in accordance with her role.

The clusters of practical standards are often, but not necessarily, shaped by the wider context in which the roles are found, such as institutional contexts. For example, teacher is a role in an institution whose goal is to educate people, and the role of teacher is shaped by the institution's policies.

But the processes by which roles become defined may vary a lot. The role of human parent is not usually determined institutionally, but by the actual needs of human children. This is a typical example of a demanding and open-ended role, because specific practical standards that must be met by the occupant of the role are set by the evolving and sometimes unexpected needs of a child, and so, are not under the direct or exclusive control of the role occupant. They are not necessarily foreseen by her when she takes up the role and they may even be difficult to come to know.

(c) The psychology of adopting and occupying roles

Psychologically speaking, what is it for an agent to adopt and occupy a practical role? A natural suggestion is that in adopting a role an agent comes to *hold herself* to the practical standards that define or constitute that role. We can begin to formulate the idea in terms of the following necessary condition:

> **Adoption and Occupation of a role (AOR):** If S adopts and occupies a role, S comes to hold herself to, and continues to hold herself to, practical standards that are constitutive of the role.

Psychologically speaking, what is it for the agent to hold herself to the standards of a role? It seems that if the agent is to hold herself to standards, she must *accept the standards as authoritative for her in her*

[129] The open-endedness of such roles and our inability to foresee the demands that they make will make on us when we adopt them is a theme of current interest (e.g. Marušić 2015, Paul 2016).

deliberation and action. But how should this idea of "accepting a standard as authoritative for one" be unpacked? We can borrow from Bratman's planning theory of intention to begin to flesh this out. (1989, 2007) When an agent accepts the standards of a role as authoritative for her, and insofar as she is rational, she experiences "rational pressure" to do the following kinds of thing:

(i) not to deliberate about ends that, she realizes, are incompatible with acting in accordance with the role;
(ii) to deliberate about means to the ends that she has because of her occupation of the role;
(iii) to act in accordance with the requirements of the role when relevant circumstances obtain.

To say a bit more, what it is for the agent to accept the standards of the role as authoritative for her is for her to experience, and often to accede to, rational pressure to accept constraints on, and demands for, role-relevant deliberation and action. The rational pressure can be understood as something that is palpable to the agent from her first-person perspective. Roughly, she has dispositions to think things such as,

(1) In a spirit of self-criticism: "I need to stop thinking about the alternatives to A-ing – I have already made up my mind to A. Moving on!"
(2) Reminding herself : "I need to find a means to A"
(3) Urging herself: "I have to A now"

Note, first, that these are not particularly odd thoughts in the mental lives of rational planning agents such as statistically normal human beings. In fact, they seem to be characteristic of everyday role-holding. The agent who adopts and occupies a demanding role, such as the role of parent, typically incurs many role-related "must" thoughts. For example, suppose that S has adopted and now occupies the role of parent. It is to be expected that S will have thoughts concerning her child, C, such as: "I must give C vitamin D", "I must find a good school in which to enrol C", and so on. These "must" thoughts do not obviously express S's desires, at least if desires are understood to involve a strong feeling of attraction or a powerful motivation: she may not feel strong attraction or motivation. Rather, when she thinks that she must do something, she seems to acknowledge the authority of a strict requirement on her to act, and to acknowledge that she would be criticisable for not so acting. These "must" thoughts offer evidence that the agent accepts the authority over her of the practical standards that, she takes it, constitute the role she occupies. They seem to be, in fact, what is involved in an agent's experiencing "rational pressure" to comply with the practical standards that constitute the role of parent.

If this is on the right track, we can add the following claim to AOR:

Holding Oneself to Practical Standards (HOPS): If S holds herself to practical standards that are constitutive of a role, she experiences, and sometimes accedes to, rational pressure to (i) avoid deliberating about role-inconsistent ends, to (ii) deliberate about role-relevant means, and to (iii) take actions required by the role.

Although a full defense of AOR and HOPS lies beyond the scope of this short paper, reflection on the everyday case of the parent suggests that they are along the right lines. But note also that we seem to rely on something like AOR and HOPS when we interpret someone as having adopted a role. Suppose that S professes to be a teacher. But suppose that she doesn't constrain her practical deliberation about ends that she knows to be inconsistent with the role and she never engages in role-relevant means-end reasoning. Suppose also that she experiences no rational pressure to engage in role-required actions, such as attending classes or preparing syllabi, and so on. She doesn't regard herself as criticisable in any way for not doing these things. She also does not accept the criticism of others if they point out, say, that she should have prepared a syllabus. When she does not seem to be responsive to pressure to do such things, it is hard to make sense of S as having adopted and as occupying the role of teacher. Perhaps, we might concede that she is a teacher "in name only" - perhaps she still has a contract to work somewhere as a

teacher. But it also seems reasonable to say that she hasn't *really* adopted the role of teacher. The plausibility of this claim can be neatly explained by appeal to the fact that the putative role-occupant doesn't seem to accept as authoritative for her the standards that are constitutive of the role - she shows no signs of feeling the rational pressure exerted by the practical standards that are constitutive of the role. This is in line with AOR and HOPS.

(d) Occupying a role VS. Committing to a role

It may still seem that an agent can occupy a role without taking the practical standards of the role to be authoritative for her. Suppose that S is bullied and threatened in an effort to force her into a marriage that she does not want. Both to avoid further pain and punishment, and because she lacks alternatives, she relents and gets married. She may accept that she must now help her spouse with certain pursuits, be faithful to her spouse, and so on, but S may feel considerable ambivalence about holding herself to such practical standards. Is it true, then, that although she occupies the role of spouse, she doesn't accept the authority of the practical standards constituting the role?

But, as argued before, if S didn't ever feel or accede to any rational pressure to constrain her practical deliberation in role-facilitating ways, or to engage in role-relevant means-end deliberation, or to act in ways demanded by the role, I think that it would be more plausible to say that although she went through the marriage ceremony, and is now socially and legally recognized as a wife, it is nevertheless true that *she hasn't really adopted* the role of spouse. It still seems plausible, I think, that adopting and occupying a role, however ambivalent one may feel about one's role, requires accepting to some extent the authority over one of the practical standards that are constitutive of the role, and as specified in HOPS, experiencing, and sometimes acceding to, certain kinds of rational pressure.

But such cases suggest at least a rough distinction between occupying a role and being committed to a role. An ambivalent or reluctant occupant of a role, S, may make some effort to meet the standards, but not much. And when she fails to meet them, she won't feel her failure deeply. In fact, she may be quite dispassionate, viewing her failures as facts that don't provoke in her much feeling of disappointment, nor need she feel much inclination to be irritated with herself for her failure. And in fact, if she feels disappointment or irritation with herself, it may be because she worries that her failure will lead, say, to divorce, and the bad consequences that she fears will follow such an eventuality. When she meets the standards non-accidentally, by her own efforts, she may feel mild relief, but not joy or delight.

How does the agent who is much more fully committed to her role differ from S? First, there is effort: she makes considerable effort to understand and meet the standards of the role. Second, there is emotional involvement: she is disposed to feel considerable disappointment and displeasure when she fails to meet the standards. Similarly, she is disposed, not just to experience, say, mild relief when she meets the standards, but is disposed to feel genuine happiness. Finally, there is valuing success in the role for its own sake: a full commitment to one's role may involve happiness at meeting the standards, not because this furthers some end that is extrinsic to the role, but just because it facilitates meeting some of the goals of the role. For example, when one meets some standard of parenthood, one is disposed to feel pleasure just because one's child is made happy (a goal of that role), and not because this success of yours will, say, mollify child services or challenge your mother-in-law's belief that you are an incompetent parent.

These points suggest the following condition on what it takes psychologically to be fully committed to a role, such as spouse or parent:

> **Full commitment to a role (FCR):** If S is fully committed to role R, (i) S is disposed to expend considerable effort in understanding and meeting the practical standards that define R; (ii) S regards meeting the practical standards of R as intrinsically valuable; (iii) S is disposed to feel considerable

disappointment with herself and with the outcome when failing to meet these standards and is disposed to feel considerable happiness with the outcome when she meets them.

There is more to be said to defend and develop FCR, and the reader will think of examples that do not fit comfortably with the (i) effort, (ii) intrinsicness or (iii) emotional investment conditions. The (ii) intrinsicness condition may seem particularly problematic: imagine the security specialist who devotes herself to ensuring that the best security solutions are found for her employer's organization and its employees in ongoing conditions of civil unrest. Must she regard meeting the practical standards of her role as intrinsically valuable if she is to be fully committed to the role? This is not obvious. As a full discussion is not possible, two remarks are in order. First, FCR fits particularly well with roles involving love of someone or something for its own sake, such as one's children, friends, nature, music, etc. If one were to concede that there are cases of FCR that do not satisfy (ii) intrinsicness, then one could attempt to restrict the roles that are covered by FCR. Second, although FCR is formulated as a necessary condition, it would be adequate in what follows to regard FCR as an informative generalization about statistically normal human beings who are deeply committed to a practical role. This is all that will be required for the argument of Section 2.

Section 2. The conflict between commitment and normative control

(a) Unthinkable practical options

Many of us have heard stories about the *tunnel vision* of addicts. Children, marriage, work, friendship, integrity, are cast aside as the addict focuses her deliberative efforts on getting a fix (e.g. Arpaly and Schroeder, 2013). There is empirical evidence indicating that the tunnel vision results from what neuroscientists picturesquely term the *hijacking* of the dopamine system by addictive substances – this results in powerful cravings, which, once triggered, make reasonable open-mindedness in practical deliberation exceedingly difficult. (Schroeder, 2010)

But non-addictive habits of thought and feeling, psychiatric disorders, taboos, feelings of disgust or rage, stereotype threat, and other things constrain practical deliberation in irrational and unreasonable ways. They give rise to "tunnel vision" – they undermine reasonable open-mindedness in practical deliberation and make practical options "unthinkable" for an agent. A "practical option" will be understood as an option for action that is within the agent's power to perform – she is able to perform that action, given current physical conditions.[130] An "unthinkable" practical option will be understood as a practical option that the agent is unable - for one reason or another - to consider in practical deliberation. (Watson 2004, 106-110)[131]

(b) Unthinkable practical options for planning and role-holding agents

But not all scenarios in which an agent is unable to consider a practical option involve irrationality or unreasonableness. Planning agents do not come to the deliberative table empty-handed, they come burdened with many prior commitments: their values, plans, policies, and their practical roles. Rationally speaking, these prior commitments should constrain practical deliberation unless the agent has a reason to doubt the reasonableness of holding on to them. Absent such doubts, and on pain of irrationality, ends incompatible with those already settled on, for example, should no longer be considered in deliberation. (Bratman 1989) Such ends become unthinkable practical options for the planning agent insofar as she is rational.[132]

---

[130] An "external option" in Smith's (2009) parlance.

[131] The literature on "is un/able to" and ability is vast and complex. I try to side-step the general issues and concern myself almost exclusively with what an agent is *rationally able* to take into consideration in deliberation, given her unrescinded commitments.

[132] We might also understand them as being "silenced" in McDowell's sense – see McDowell (1979).

Practical roles, such as that of spouse or parent, can also quite radically but rationally constrain the agent's practical deliberation. Consider "Caregiving Cases" – cases in which an agent has (a) a very onerous practical role to which (b) she is deeply committed, thereby satisfying FCR. The single working parent of twin toddlers who has little outside help, or the devoted wife of a spouse with dementia, also without much outside help, are examples of Caregiving Cases. Their plates are very full. In the terms of Section 1, they accept the authority over them of the practical standards constituting their caregiving role. Because of the onerousness of the role in Caregiving Cases, and in line with Adopting and Occupying a Role (AOR), their attention is drawn repeatedly to the practical standards that they hold themselves to, given their role. In line with Holding Oneself to Practical Standards (HOPS), their mental lives are, in fact, dominated by role-related "must" thoughts. And when they engage in deliberation, it is usually deliberation that allows them to crystallize and meet the practical standards of their complex, wide-ranging, and open-ended practical roles.

We might think of the kind of practical deliberation that dominates their deliberative lives as *technocratic deliberation*. It is "technocratic" insofar as it is tightly organized around meeting the standards of their practical role. It is not concerned with the meaningfulness or value of that role, nor is it concerned with the importance of other values that the agent has.

Nevertheless, let's suppose that the agents in our Caregiving Cases value highly further education, artistic expression, friendship, travel, and so on. Let's also suppose that these ends, if adopted, would prevent them from fulfilling their caregiving roles. Given this, it seems that they will be rationally deterred even from deliberating about these alternative ends, as long as they remain committed to their caregiving roles – surely such deliberation is itself a threat to their commitment. But if this is correct, it is not just that the agents don't have the time or motivation to deliberate about alternative ends, it is that such ends – such practical options - have become rationally "unthinkable" for them. But isn't this a loss? Let's try to be a bit more precise about this.

(c) Role-commitments and unthinkable practical options – more precisely

First, let's assume that the purpose of practical deliberation about ends is to settle the question of whether to pursue those ends. Practical deliberation "characteristically yields … a judgement – which has an internal, necessary relation to subsequent action or intention." (Stroud 2003, 122) If this is right, practical deliberation is very different from practical musing, say – its purpose when it concerns competing ends is to arrive at a decision about what end to adopt and act on. It has an internal relation to action and is not simply a matter of considering which ends might be appealing or permissible, and so on.

Second, our caregiving agents have adopted the role of caregiver, which, if Adopting and Occupying a Role (AOR) and Holding Oneself to Practical Standards (HOPS) are on the right track, involves *accepting* the practical standards that define the role as authoritative for them in deliberation and action.

Third, let's suppose that the agents know that a practical option that they value highly – such as pursuing a PhD, or travelling the world - would, if it were adopted by them, make it impossible for them to fulfil their caregiving role in the manner that they have been doing to this point. Let's call this kind of practical option/end a "highly Valued but Conflicting (with the fulfilment of the caregiving role) End" – a VCE.

Fourth, it seems that deliberation about a VCE necessitates bracketing the authority of a caregiving role: in deliberating about whether to adopt the VCE, agents in Caregiving Cases must, rationally speaking, already call into question whether they will continue to hold themselves to the practical standards of the caregiving role. Calling this into question is to no longer *accept* the authority of the caregiving role – it is to bracket it in order to take seriously the possibility of adopting a practical option that is inconsistent with holding the caregiving role.

As long as the agent remains committed to her caregiving role, it looks like deliberating about VCEs is rationally ruled out. Although these practical options or ends are highly valued by the agent, they are rationally unthinkable practical options.

To put things another way, there seem to be three options open to the agent in a Caregiving Case who wishes to honour a VCE, and none of these is palatable. First, she might engage in practical musing that is not really practical deliberation. This may leave the authority of the caregiving role in place, but it does not take the VCE very seriously. The VCE is not allowed to shape thought that has an "internal relation to intention and action". The second option is to engage in irrational practical deliberation by refusing to acknowledge the conflict between the caregiving role and the VCE. It may leave the authority of the current role more or less intact, but it fails to reflect the facts, and so, is very defective as deliberation. The third option is that the agent's deliberation can reflect the facts about the onerousness of her current caregiving role and its conflict with the VCE. But to deliberate, she must bracket the authority of the practical standards of the caregiving role.

The last of the three options is preferable because it is the most reasonable and rational option that also takes the VCE seriously. But it involves at least temporarily suspending the acceptance of the authority of the caregiving role. For the agent who satisfies FCR, this is a deeply troubling, if not repugnant, option. For her, suspending her commitment is likely to be unacceptable even when she suspends in the knowledge that her end is likely to be reinstated after deliberation. If she satisfies FCR, she has expended great effort in promoting the well-being of her charge, she is emotionally invested in this venture, and values the well-being of her charge for its own sake. Although it does not seem irrational or unreasonable of her to value highly a life with serious and absorbing intellectual pursuits, or deep friendships, or the joys of travel, say, these VCEs seem to be unthinkable practical options.

(d) Unthinkable practical options and loss of normative control

Not realizing valued ends is a loss. But beyond this somewhat ordinary loss, there is the loss of not having the opportunity to set these ends aside in a process of careful deliberation about the relative merits of competing ends. Let's consider such a process and what value it may have to the agent.

Everyday practical deliberation about ends that we value highly is often not a linear or simple process. The agent's thought processes may be structured only by the felt pressure to reach a conclusion about which of two competing ends to adopt, but be otherwise wide-ranging. It may involve serious, effortful, and sustained consideration of the competing ends, while taking into account, and re-evaluating, prior plans, policies, and values. It may involve the collection and assessment of evidence, and the review, rejection, formation, or withholding of beliefs. This process may take a lot of time (and sleepless nights). And it may have emotional and imaginative components - it may involve conative and cognitive states, and states with phenomenal aspects. When one deliberates, for example, about whether to leave one's friends and family to pursue a distant work project, one may reflect on and experience the love that one feels for one's family members, or one may experience the joy and excitement that one feels at the prospect of tackling a professional challenge. These phenomenal states shape the deliberative process. The process may also involve ruminating on philosophical questions such as how important to a meaningful life professional development is, and whether one has certain moral obligations, and so on. All in all, this may be an arduous intellectual and emotional process.

But it can also be rewarding. Aside from the value of the intellectual exploration it involves, other goods become available if a clear conclusion is reached. If the conclusion is backed by what is, and what the agent judges is, a careful, thorough, and reasonably open-minded deliberative process, this will allow the agent to confidently reject as insufficiently important or desirable the end that she decides against adopting. Although such a conclusion may bring sadness about what is being given up, a clear conclusion may still allow her to embrace the "winning" valued end with greater peace of mind and a greater

readiness for wholehearted commitment. In addition, an agent may hold some of the deliberative process in memory, so that she may review and re-affirm her conclusion at a later time. Or indeed, should she re-deliberate, she does not have to start from scratch, but can call on the fruits of her earlier deliberation to re-start the process. And of course, the agent is able to explain her choice to others and be a full participant in social practices of self-justification.

For comparison, consider a case where an end, E_1 – a highly-valued end, that *would have won out* in a robust deliberative process – had been thrust upon an agent, S, by circumstances beyond her control. Suppose also that a VCE, E_2, had also been put decisively beyond S's reach by such circumstances. Even though it may be true that S would have chosen E_1 and rejected E_2 had she had the opportunity to deliberate, without undergoing the deliberation she may lack confidence or enthusiasm about her pursuit of E_1, and she may continue to worry that E_2 would have been the better one to pursue.

The foregoing suggests that a thorough practical deliberation about competing ends is valuable beyond its role in allowing agents to choose among those ends. Such deliberation can provide intellectual and emotional closure that aids in the wholehearted pursuit of an end and assuages the agent's worries about giving up a VCE.

A thorough, careful, conclusive practical deliberation about ends is plausibly understood as an exercise of normative control: it is, after all, a process that allows the agent to arrive at intentions (and hopefully actions) that are in line with what she most values. It is, in fact, plausible to suppose that it is a key element in self-government.

If the conclusion of subsection (c) is correct, such a deliberative process is not rationally available to agents of our Caregiving Cases who remain highly committed to their caregiving role. Exercising such normative control puts unreasonable demands on such agents to step aside from a role that they deeply love and care about. Even if they may truly say "I can do no other" and are usually content with their caregiving role, and even if they do not pine daily for some VCE, insofar as they do cherish some VCE, but do not have the benefits of a free and open practical deliberation, they are missing out. They are unable to gain deep understanding of the merits of their choice, self-understanding with respect to the choice they have made, reassuring memories of a deliberative process well executed, and the ability to articulate to others why they have rejected the VCE.

(e) Further discussion

It is hopefully clear that the Caregiving Cases are not just problematic for the "ordinary" reason that difficult circumstances sometimes prevent the realization of valued ends. We all give up on valued ends in the face of conflicting ends or obligations, and of course, we also fail in our attempts to execute our plans, and must abandon valued ends as a result. But in ordinary cases where we sacrifice the pursuit of something we value, we may reach the conclusion to abandon by careful practical deliberation. We exercise our capacities for normative control, and this process may result in the conviction that our decision is the best available. These ordinary cases do not involve deliberative "silencing", however painful they may otherwise be.

In fact, Caregiver Cases are closer to the case of the addict who is unable to take seriously in a reasonably open-minded deliberation things that she continues to value. In the grip of craving, the addict suffers from tunnel vision, a lack of sensitivity to what she values. In later sobriety, she may regret that she simply turned her back on these things without taking a deliberative stand on the abandoned ends. In spite of important differences, both the addict and the caregiver are unable to exercise the kind of normative control that may bring self-understanding and reassurance that they have done, and are doing, the right thing.

It might be objected that the *real* trouble in the Caregiving Cases is not onerous role-based commitments, but a social world that does not provide a safety net when things go wrong for agents with such commitments - divorce, dementia, and other things can make role-based commitments borderline unmanageable and lock the agent into relentless technocratic deliberation. The objector points to an important connection to social settings, which should be explored if we are to better understand the value of commitment in general. Nevertheless, this objection ignores the fact that commitment is a key problematic element in the cases considered. Of course, more favourable circumstances will ease the conflict between caregiving roles and VCEs, but it still seems that commitment is intrinsically problematic: by its nature it has the potential to silence deliberation and hamper a certain kind of normative control.

It should be emphasized that the aim of the discussion is not to take on the very large question of whether commitment is, in general, valuable. I also do not take a stand on when (or if) it is rational or reasonable to act in accordance with a prior commitment. My more limited aim is, rather, to open questions that are often left to one side: what is lost to the committed agent (and the planning agent, more generally) when she commits? If I am right and she must forswear practical deliberation about VCEs, what is lost? If we accept that (i) practical deliberation about ends is an exercise of normative control and (ii) that practical deliberation about competing ends can generate valuable self-understanding and practical reassurance, then there are serious potential costs to committing.

I have said that a VCE cannot be assessed by practical musing. But it might be objected that an agent can, in fact "honour" a VCE in this way. An agent in a Caregiving Case may quite deeply reflect on a VCE, such as pursuing a PhD in art history, by imagining what a life devoted to this would be like. This imaginative process will allow her to explore the pleasures and pitfalls of such a life without threatening her commitment. This objection raises important questions about the nature of practical deliberation and the nature of imagination, and the relationship between them. Accordingly, a full response goes well beyond me here and a brief suggestion will have to suffice. If the agent's imagining is not embedded in a practical deliberation, then it is not clear that it can answer the question of whether or why she should forswear pursuing a PhD in spite of its deep appeal. In being an exercise in imagination only, it does not involve weighing a life with a PhD against a life of caregiving, coming to appreciate the merits and demerits of each, for the sake of arriving at a conclusion about which end to adopt. Such a process may allow her to decisively set aside the PhD, say, and this may give her self-understanding and the reassurance that she is doing the right thing. But an imaginative exercise, however rich, does not seem to do this. And if it did, then the imaginative exercise would seem to have deliberative elements, which, for the agent who is rational, will encroach on her caregiving commitment, and it is plausible to worry that we will again face the problem discussed in subsection (c) above: how can an agent rationally engage in this kind of imaginative process about a VCE when she is already highly committed? If this is along the right lines, there is a legitimate worry that this imagination objection faces a dilemma: either it does not help the agent much in restoring her normative control and providing her with the benefits comparable to those resulting from a thorough practical deliberation, or it does help her, but this help is in deep rational tension with her caregiving commitment.

Concluding Remarks

In Drowning we may find it appealing that in the heat of a terrible moment the spouse acts reliably out of love – in doing so, she reveals her deep commitment to the other and is not assailed by chilly thoughts of what is right or wrong. But this reliable response seems to manifest a technocratic facet of rational practical agency – a facet devoted to producing decisions and actions that are required of her *given her commitment*, but not devoted to thorny questions about whether and when to commit, or what the value of committing may be. Such responses seem to involve a loss of valuable normative control. And this is

arguably reflected in the spouse's feeling somewhat dumbfounded by, and uncomfortable with, her (i) volitional incapacity and (ii) deliberative silence. I have not argued that the spouse on the shore *should coolly reason* about who to save. I do not know what she should do. What I have tried to establish is that we have reasons to think twice before either singing the praises of commitment or before looking askance at the urge to deliberate.

Works Cited

Arpaly, N. and Schroeder, T. 2013. 'Addiction and Blameworthiness', in Levy, N. ed. *Addiction and Self-Control*, Oxford University Press. 214-238

Bratman, M. 1989. *Intention, Plans, and Practical Reasoning*. Stanford: CSLI Publications.

------ 2007 *Structures of Agency*. New York: Oxford University Press.

Chang, R. 2013 'Commitments, Reason, and the Will' in Russ Shafer-Landau (ed.), *Oxford Studies in Metaethics: Volume 8*. Oxford University Press. pp. 74-113.

Ferrero, L. 2010. 'Decisions, Diachronic Autonomy, and the Division of Deliberative Labor', *Philosophers' Imprint*, 10/2: 1–23.

Frankfurt, H. 1988. *The Importance of What We Care About*. Cambridge University Press.

Kennett, J. 2013. 'Just Say No? Addiction and the Elements of Self-Control' in Levy, N. ed. *Addiction and Self-Control*, Oxford University Press, 144-164.

McDowell, J. 1979. Virtue and Reason, *The Monist*, 62:3, 331-350.

Marušić, B. 2015. *Evidence and Agency*, Oxford University Press.

Paul, L. 2016. *Transformative Experience*, Oxford University Press.

Railton, P. 2006. 'Normative Guidance' in *Oxford Studies in Meta-Ethics*, Vol. 1, Schafer-Landau, R. (ed.) Oxford University Press, 3-33.

------ 2009, 'Practical Competence and Fluent Agency', *Reasons for Action*, Cambridge University Press, Sobel, D. and Wall, S. (eds.) 81-115.

Schroeder, T. 2010. 'Irrational Action and Addiction' in Ross, Kincaid, Spurrett, and Collins, eds. *What is addiction?*, MIT Press, 391-407.

Shpall, S. 2014. 'Moral and Rational Commitment', *Philosophy and Phenomenological Research*, 88:1.

Smith, M.N. 2009. 'Practical Imagination', *Philosopher's Imprint* 10:3.

Stroud, S. 2003. 'Weakness of will and practical judgment' in Stroud, S. and Tappolet, C. (eds.) *Weakness of Will and Practical Irrationality*, Oxford University Press.

Watson, G. 2004. *Agency and Answerability*, Oxford University Press.

Williams, B. 1981. "Persons, Character, and Morality." In *Moral Luck*, 1-19. Cambridge: Cambridge University Press.

SELF-MASTERY IN PLATO'S *LAWS*

Brian Reese

In the first book of Plato's *Laws*, an unnamed Athenian seeks to clarify the notion of *self-mastery* (τὸ κρείττω ἑαυτοῦ). We humans are like puppets, he says, pulled in opposing directions by various 'strings' or 'cords' within us. We are pulled toward vice by 'iron' cords associated with feelings of pleasure and pain. We are pulled toward virtue by a 'golden' cord associated with reasoning and law. According to the most natural reading of this stretch of text, self-mastery consists in the victory of reason in its struggle against the opposing pulls of pleasure and pain. The victory of reason is said to issue in *virtue*.

This claim about virtue, however, is surprising. Both before and after the so-called 'puppets passage' (644d7–645b1), the Athenian maintains that virtue consists in the harmonious relations among potentially conflicting parties. Indeed, he explicitly claims that it is best when such parties are reconciled to one another, not when one party forcibly subordinates another in conflict (628c9–e1). So which account is to be preferred? Is virtue ultimately to be understood as the *victory* of reason over feelings of pleasure and pain, or as the *harmonious* relations among them? While recent commentators have produced and defended a number of interpretations aiming to resolve this dilemma about virtue, the concept of self-mastery is very often sidelined in the process.[133] This, I will argue, is a mistake: a proper understanding of self-mastery is essential for any fully satisfying resolution to this dilemma about virtue.

The plan for the paper is as follows. After examining the passages where self-mastery is foregrounded in the first book of the *Laws* (§1), I discuss the puppets passage and the attending dilemma about virtue that has exercised recent commentators (§2). I then provide a novel resolution to it (§3). I argue that the Athenian is operating with two distinct conceptions of self-mastery in view, which can be appreciated only after uncovering an overlooked distinction between two types of *moderation* (σωφροσύνη). Armed with these distinctions, it is no longer necessary to choose between the two seemingly conflicting accounts of virtue canvassed above (§4). A careful analysis of self-mastery dissolves the dilemma by revealing how both accounts can represent legitimate ways of thinking about virtue.

§1 SELF-MASTERY & CONFLICT

The puppets passage is not the first place in the *Laws* where self-mastery is discussed. Indeed, it is striking that the work begins with a sustained and careful treatment of self-mastery, which then structures much of the ensuing conversation (627a3). No sooner have an unnamed Athenian and his two interlocutors, Clinias and Megillus, agreed to examine "constitutions and laws" (625a6) than they seek to identify the primary aim of the legislator.[134] Clinias proposes that it is to ensure victory in war (626a4). This initial proposal is then subjected to sustained examination. When Clinias is asked to specify the scope of the war he has mind, he is first led by the Athenian to acknowledge—and then to enthusiastically maintain—that wars exist not just between states, but also between villages within states, households within villages, and individuals within households (626c3). The most pervasive sort of conflict, however, is discovered to be the one waged within each of us. Each person, we learn, "is pitted against themselves" (626d8). The victory one achieves in such cases is agreed to be the "first and best (πρώτη τε καὶ ἀρίστη)" of all victories (626e2).

[133] Annas (1999), Belfiore (1986), Bobonich (2002), Frede (2010), Meyer (2015 and 2018), Schofield (2016), Wilburn (2012)
[134] Translations from Books I-IV of the *Laws* by Meyer. Translations from Books V-XII of the *Laws* by Schofield (2016) and sometimes Saunders (1970). Translations occasionally altered for consistency with terminology in the present paper.

This is the sort of struggle that will later be taken up in the puppets passage.[135] Initially, however, no consideration is given to the parties involved in what I will henceforth call *intrapersonal* cases of conflict. Instead, the Athenian and his interlocutors focus on some *interpersonal* cases—those of familial and political conflict.

A crucial feature of such conflicts, the interlocutors agree, is that they are waged between a better and a worse party.[136] Self-mastery is thus initially glossed as the victory of the better party over the worse, while self-defeat is glossed as the victory of the worse party over the better (627b2).[137] Now, although it may initially seem strange to use the locutions 'self-mastery' and 'self-defeat' to describe cases of interpersonal conflict, the Athenian makes it clear that he is unconcerned with common linguistic usage (627d3–5). It is the underlying phenomenon that he is most interested in, and his general strategy seems to be to extend the notion of self-mastery beyond its principle application to individuals (626e–628e). The Athenian deploys an example in order to help illustrate the expanded meaning of this term: in a family comprising several brothers—the majority of whom we are to suppose are unjust—the family as whole would rightly be called 'self-defeated' if the unjust brothers prevail in conflict, and 'self-mastered' if the just minority prevail. So too for the state: "whenever the better people are victorious over the inferior masses, the city would correctly be called self-mastered" (627a7).[138]

In all such cases of conflict, the Athenian maintains that it would be far better to reconcile the conflicting parties than to forcibly subordinate the inferior party to the superior one:

> The best is neither war nor faction (one should pray to be spared the necessity of either) but rather peace and friendship. Victory of a city over itself, it would seem, is not best but a necessity (οὐκ ἦν τῶν ἀρίστων ἀλλὰ τῶν ἀναγκαίων). To think otherwise is like supposing that a disease-ridden body is performing at its best after being flushed out by a purgative—with no thought to the case of a body that needs no such treatment. For the same reason, no proper statesman will assess the happiness of either a city or an individual (ἢ καὶ ἰδιώτου) solely and primarily with a view to war against external enemies, and no lawgiver is any good unless they regulate military matters for the sake of peace, rather than regulating peacetime for the sake of war. (628c9–e1)[139]

Far better than a recovering body is one that was never sick, and far better than a state recovering from civil war is one in which citizens have continually enjoyed peace and friendship.[140] Indeed, peace and friendship

---

[135] Clinias initially speaks of achieving "victory over oneself" (τὸ νικᾶν αὐτὸν ἑαυτὸν). However, the Athenian immediately reformulates this as follows: "each of us, a single individual, is either master of or defeated by himself (ὁ μὲν κρείττων αὑτοῦ, ὁ δὲ ἥττων)" (626e8). Clinias is made to both repeat and explicitly accept the Athenian's reformulation (627a3).

[136] Note that this excludes the possibility of conflict between equal parties.

[137] Cf. *Republic* 430e7

[138] ὀρθῶς ἂν αὖτη κρείττων τε ἑαυτῆς λέγοιθ' ἡ πόλις.

[139] Τό γε μὴν ἄριστον οὔτε ὁ πόλεμος οὔτε ἡ στάσις, ἀπευκτὸν δὲ τὸ δεηθῆναι τούτων, εἰρήνη δὲ πρὸς ἀλλήλους ἅμα καὶ φιλοφροσύνη, καὶ δὴ καὶ τὸ νικᾶν, ὡς ἔοικεν, αὐτὴν αὑτὴν πόλιν οὐκ ἦν τῶν ἀρίστων ἀλλὰ τῶν ἀναγκαίων· ὅμοιον ὡς εἰ κάμνον σῶμα ἰατρικῆς καθάρσεως τυχὸν ἡγοῖτό τις ἄριστα πράττειν τότε, τῷ δὲ μηδὲ τὸ παράπαν δεηθέντι σώματι μηδὲ προσέχοι τὸν νοῦν, ὡσαύτως δὲ καὶ πρὸς πόλεως εὐδαιμονίαν ἢ καὶ ἰδιώτου διανοούμενος οὕτω τις οὔτ' ἄν ποτε πολιτικὸς γένοιτο ὀρθῶς, πρὸς τὰ ἔξωθεν πολεμικὰ ἀποβλέπων μόνον καὶ πρῶτον, οὔτ' ἂν νομοθέτης ἀκριβής, εἰ μὴ χάριν εἰρήνης τὰ πολέμου νομοθετοῖ μᾶλλον ἢ τῶν πολεμικῶν ἕνεκα τὰ τῆς εἰρήνης.

[140] It should be noted that the Athenian's endorsement of friendship over faction coheres well with the account in the *Republic*. There, psychic harmony is likened both to a city that is free of conflict (441e9–442d3), and to a body that is free of disease (444c3–445b4). These, of course, are the very same analogies that the Athenian deploys.

seem to be preferred to faction—not only in the state, but also in the individual.[141] It is only when we arrive at the puppets passage, however, that the Athenian is finally prepared to offer an analysis of the parties involved in the latter sort of conflict. It is to this passage that we now turn.

§2 THE PUPPETS PASSAGE

The Athenian first undertakes an analysis of the parties involved in cases of intrapersonal conflict toward the end of the first book of the *Laws*. Reminding his interlocutors of their earlier discussion, the Athenian now wishes to clarify his thoughts on self-mastery in the intrapersonal case by means of an illustration (644c3). We are each a single individual, he says, but we have within us various forces that pull in opposing directions, like so many 'strings' or 'cords' (644c5). We are in this respect like puppets:

> Consider each of us, living beings that we are, to be a puppet of the gods—whether constituted as the gods' plaything, or for some serious purpose, we have no idea. What we do know is that these passions in us (τὰ πάθη ἐν ἡμῖν) are like cords or strings that tug at us and oppose each other. They pull against each other (ἀνθέλκουσιν) towards opposing actions (ἐναντίας πράξεις) across the field where virtue is marked off from vice. Our account singles out one of these pulls (ἕλξεων) and says that each of us must follow it and pull against (ἀνθέλκειν) the other cords, never loosening our grip on it. This is the sacred and golden guidance (ἀγωγὴν) of reasoning (λογισμοῦ), also called the city's common law [...] One must always pitch in (συλλαμβάνειν) with the noblest guidance, that of law, since reasoning—although it is noble—is gentle rather than violent, so its guidance requires helpers if our golden element is to be victorious (νικᾷ) over the other cords. (644d7–645b1)[142]

We are pulled toward vice by 'iron' cords associated with feelings (πάθη) of pleasure and pain, and we are pulled toward virtue by a 'golden' cord associated with reasoning (λογισμός) and law (644c7).[143] If the golden cord is to emerge victorious in this tug-of-war battle across the field where virtue is marked off from vice, then it is necessary for us to 'pitch in' with it. The Athenian now concludes:

---

[141] While the Athenian indicates that his conclusion applies to *both* an individual and a city (628d4), many commentators are quick to point out that he fails to state explicitly that this means the individual must have internally harmonious relations.

[142] Περὶ δὴ τούτων διανοηθῶμεν οὑτωσί. θαῦμα μὲν ἕκαστον ἡμῶν ἡγησώμεθα τῶν ζῴων θεῖον, εἴτε ὡς παίγνιον ἐκείνων εἴτε ὡς σπουδῇ τινι συνεστηκός· οὐ γὰρ δὴ τοῦτό γε γιγνώσκομεν, τόδε δὲ ἴσμεν, ὅτι ταῦτα τὰ πάθη ἐν ἡμῖν οἷον νεῦρα ἢ σμήρινθοί τινες ἐνοῦσαι σπῶσίν τε ἡμᾶς καὶ ἀλλήλαις ἀνθέλκουσιν ἐναντίαι οὖσαι ἐπ' ἐναντίας πράξεις, οὗ δὴ διωρισμένη ἀρετὴ καὶ κακία κεῖται. μιᾷ γάρ φησιν ὁ λόγος δεῖν τῶν ἕλξεων συνεπόμενον ἀεὶ καὶ μηδαμῇ ἀπολειπόμενον ἐκείνης, ἀνθέλκειν τοῖς ἄλλοις νεύροις ἕκαστον, ταύτην δ' εἶναι τὴν τοῦ λογισμοῦ ἀγωγὴν χρυσῆν καὶ ἱεράν, τῆς πόλεως κοινὸν νόμον ἐπικαλουμένην [...] δεῖν δὴ τῇ καλλίστῃ ἀγωγῇ τῇ τοῦ νόμου ἀεὶ συλλαμβάνειν· ἅτε γὰρ τοῦ λογισμοῦ καλοῦ μὲν ὄντος, πρᾴου δὲ καὶ οὐ βιαίου, δεῖσθαι ὑπηρετῶν αὐτοῦ τὴν ἀγωγήν, ὅπως ἂν ἐν ἡμῖν τὸ χρυσοῦν γένος νικᾷ τὰ ἄλλα γένη.

[143] Cf. R*epublic* 611c3

> Here is how we may vindicate this tale of virtue (μῦθος ἀρετῆς) that likens us to puppets. It makes clearer, in a way (τρόπον τινὰ), what is meant by 'self-mastery' and 'self-defeat' (τὸ κρείττω ἑαυτοῦ καὶ ἥττω εἶναι), as well as the manner in which a city and an individual ought to live. (645b1)[144]

Given that the point of this illustration is to help clarify the meaning of 'self-mastery' and that the Athenian has now identified the salient parties in cases of intrapersonal conflict, he would finally seem to be in a position to assert—just as he had when discussing the interpersonal cases—that a condition of *harmony* or friendship is to be greatly preferred to the mere *victory* of the golden cord in its struggle against the iron cords. But surprisingly, the Athenian makes no such claim.[145]

Has he simply forgotten his earlier assertion that victory over obstinate forces is to be strongly dis-preferred to reconciliation and harmony (628d1)? Or is the Athenian now very consciously highlighting an important dis-analogy between interpersonal relations on the one hand (where harmony is the goal) and intrapersonal relations on the other hand (where victory is the goal)? The most straightforward reading of the puppets passage seems to suggest that the Athenian construes virtue in the intrapersonal case—for he calls this a tale of virtue—as the victory of the golden cord over the iron cords. That there is absolutely no mention here of a better condition involving harmony or friendship seems strange—and cries out for explanation.[146]

2.1 Conflict & Harmony: Interpretive Horns

Many commentators have tried to make sense of the rather conspicuous absence of any mention of harmony or friendship in the puppets passage. I here canvas the two most prominent interpretive strategies. The first (and perhaps most straightforward) way to explain this absence is to simply take the passage at its word: virtue in the intrapersonal case does not preclude conflict in the way that it does in the interpersonal cases.[147] On this straightforward reading, virtue as described in the puppets passage consists in the victory of the golden cord over the iron cords, just as the Athenian seems to assert. I will label this the CONFLICT model of virtue, as virtue is to be understood as the victory of a better party over a worse in cases of conflict.[148]

One obvious problem for those who read the puppets passage as endorsing the CONFLICT model of virtue, however, is that it is hard to square with the account of virtue endorsed by the Athenian in the interpersonal cases, where harmony had been greatly preferred to victory. It is also hard to square with the account of intrapersonal virtue that the Athenian soon recommends. In the very opening lines of second

---

[144] καὶ οὕτω δὴ περὶ θαυμάτων ὡς ὄντων ἡμῶν ὁ μῦθος ἀρετῆς σεσωμένος ἂν εἴη, καὶ τὸ κρείττω ἑαυτοῦ καὶ ἥττω εἶναι τρόπον τινὰ φανερὸν ἂν γίγνοιτο μᾶλλον ὃ νοεῖ, καὶ ὅτι πόλιν καὶ ἰδιώτην, τὸν μὲν λόγον ἀληθῆ λαβόντα ἐν ἑαυτῷ περὶ τῶν ἕλξεων τούτων, τούτῳ ἑπόμενον δεῖ ζῆν

[145] Cf. *Laws* 803c5 and 804b3, where talk of 'puppets' is resumed.

[146] It is not just strange given what has been said here in the *Laws*, but also given the more familiar account in the *Republic*.

[147] Both Belfiore (1986, pp. 428–433, 429) and Bobonich (2002, pp. 289, 350, 546n122) have advanced such interpretations.

[148] In labeling this the CONFLICT model, I am following Meyer (2018).

book of the *Laws*, he endorses an account of intrapersonal virtue that explicitly takes harmony as the goal:

> If pleasure and liking and pain and hatred develop correctly in our souls when we are not yet able to grasp the account (τὸν λόγον), and when we do grasp the account they harmonize (συμφωνήσωσι) with it because they have been correctly trained by the appropriate habits, this harmony is virtue in its entirety (ἡ συμφωνία σύμπασα μὲν ἀρετή). (653b2)[149]

The Athenian maintains this account of intrapersonal virtue for the duration of the *Laws*. So those who read the puppets passage as straightforwardly endorsing an account of virtue according to which a better party subordinates an inferior one have some explaining to do. First, why would the Athenian endorse this sort of account in the puppets passage when he had *already* endorsed harmony over victory in the interpersonal cases?[150] Second, why would the Athenian endorse this sort of account in the puppets passage only to cast it aside in the opening lines of the very next book?

Another common interpretive strategy is to simply deny that the sort of virtue evinced by the victory of the golden cord over the iron cords is really an instance of conflict at all. Commentators who advance this sort of reading offer interpretations according to which the puppets passage is not describing a case of conflict amongst the cords, but rather a case of harmony.[151] I will label this competing account the HARMONY model of virtue, as virtue is to be understood as the harmonious relations between better and worse parties.[152] On this sort of interpretation, the Athenian and his interlocutors have already agreed, by the time we reach the puppets passage, that the HARMONY model of virtue is, in all cases, superior to the CONFLICT model of virtue.[153] The appeal of this sort of interpretation is that it faces none of the difficulties presented by the more straightforward reading of the puppets passage described above, since it maintains that the Athenian consistently endorses the HARMONY model of virtue throughout the *Laws*.[154]

The obvious problem for this interpretative strategy, however, is that it must somehow explain away the many explicit references to conflict in the puppets passage—where pleasure, pain and reasoning are all described as distinct 'pulls' (ἕλξεων) that "draw against each other (ἀνθέλκουσιν) towards opposing actions (ἐναντίας πράξεις)" (644e3).[155] The golden cord emerges victorious (νικᾷ), it would seem, not by harmonizing or reconciling with the other cords, but by overpowering them. Indeed, the passage explicitly "singles out one of these pulls (ἕλξεων)" as the golden cord and says that each of us must help it "pull against (ἀνθέλκειν) the other cords." Now, if such textual hurdles are not simply insurmountable, they at least stretch the plausibility of this interpretative strategy.

---

[149] ἡδονὴ δὴ καὶ φιλία καὶ λύπη καὶ μῖσος ἂν ὀρθῶς ἐν ψυχαῖς ἐγγίγνωνται μήπω δυναμένων λόγῳ λαμβάνειν, λαβόντων δὲ τὸν λόγον, συμφωνήσωσι τῷ λόγῳ ὀρθῶς εἰθίσθαι ὑπὸ τῶν προσηκόντων ἐθῶν, αὕτη 'σθ' ἡ συμφωνία σύμπασα μὲν ἀρετή

[150] Arguably, he had also endorsed harmony over victory in the individual case. (See n9 above; cf. *Republic* 409b4–e1)

[151] We may suppose that this earlier endorsement is what the Athenian is referring to when he asks his interlocutors to recall their previous agreement that those who rule themselves are good at 644b6.

[152] In labeling this the HARMONY model, I am again following Meyer (2018).

[153] The Athenian, paraphrasing Clinias' remarks, glosses 'self-mastered' as 'good' at 627b7–8.

[154] Annas (1999, pp. 142–44), Frede (2010, pp. 217–20), Schofield (2016, pp. 146–48) and Wilburn (2012, pp. 29–35) have all advanced such interpretations.

[155] Cf. Meyer (2018). The verb ἀνθέλκειν is repeated at 644e6. The pull of reason is also referred to as 'guidance' (ἀγωγήν) at 645a1 and 645a7, which also suggests that it is a sort of 'pull.' Plato often uses this word, for instance, when describing the non-rational pull of appetites and emotions (cf. *Republic* 604b1 and *Phaedrus* 238c3).

2.2 Conflict & Harmony: Splitting the Horns

So there remain serious difficulties for those who opt for either interpretive horn. But there remains a third interpretive strategy that seeks to split them. According to this third interpretative strategy, when the Athenian takes up and develops Clinias' initial suggestion that victory in war is the aim of the legislator (626a4), he is developing an account of virtue (one modeled on conflict) that he does not himself endorse—but which he knows he can use as a stepping-stone to arrive at his own more fully developed account of virtue (one modeled on harmony).[156] Such an interpretation not only has the benefit of being able to read the puppets passage as straightforwardly endorsing the CONFLICT model of virtue, but it can do so without thereby committing the Athenian to it. It is thus able to avoid the difficulties faced by the previous two interpretive strategies, while preserving what is most plausible in each.

Unfortunately, this interpretative strategy is not without its own set of difficulties. According to the view it advances, on the most natural reading of the stretch of text we are concerned with, the Athenian and his two interlocutors jointly accept the following three claims:

(a) Virtue is to be understood as self-rule (ἄρχειν αὐτῶν)[157]

(b) Self-rule is to be understood as self-mastery (τὸ κρείττω ἑαυτοῦ)

(c) Self-mastery is to be understood as the victory of the golden cord over the iron cords

The problem is that these three claims collectively entail a commitment to the CONFLICT model of virtue.[158] In order to appreciate why this is such a problem, consider the sort of argument that this interpretive strategy offers to motivate the now-familiar dilemma about virtue:

(1) If virtue is to be understood as the victory of the golden cord over the iron cords in the puppets passage, then virtue entails conflict
(2) If virtue entails conflict, then the puppets passage supports the CONFLICT model of virtue
(3) Virtue is to be understood as the victory of the golden cord over the iron cords in the puppets passage (by a–c)
(4) Therefore, virtue entails conflict (by 1 and 3)
(5) Therefore, the puppets passage supports the CONFLICT model of virtue (by 2 and 4)
(6) CONFLICT is inconsistent with HARMONY as an account or model of virtue (implicit)
(7) If the Athenian endorses HARMONY, then he cannot consistently endorse CONFLICT (by 6)
(8) The Athenian endorses HARMONY

[156] Meyer (2018), p. 108
[157] *Laws* 645b2; Meyer (2018), p. 98
[158] Meyer (2018), pp. 99, 107–8

(9) Therefore, the Athenian cannot consistently endorse CONFLICT (by 6, 7 and 8)

This argument formalizes the dilemma: since the CONFLICT and HARMONY models of virtue are inconsistent with one another, the Athenian can only endorse one of them. This interpretation maintains that the Athenian endorses the HARMONY model of virtue. But it also grants that the Athenian endorses (a)–(c), which commit him to the CONFLICT model of virtue. Since the Athenian appears to endorse both models of virtue, this interpretation is forced to conclude that there is simply "no satisfactory resolution" to the dilemma.[159] The best it can do is psychologize: even granted that the Athenian explicitly endorses claims that commit him to the CONFLICT model of virtue, it is not the case that he *really* endorses these claims. He only pretends to in order to prepare his interlocutors to eventually accept the HARMONY model of virtue.

2.3 Toward a Satisfactory Resolution

At this juncture, it is perhaps worth reemphasizing that the stated aim of the puppets passage is to help clarify the notion of self-mastery (644c1, 645b2), *not* virtue.[160] The near-exclusive focus on virtue by commentators is therefore somewhat unfortunate. Such a focus has meant, more often than not, that self-mastery is either downplayed or altogether sidelined in their accounts. This, I will now argue, is a mistake.

By focusing on self-mastery rather than virtue, I uncover an overlooked distinction that the Athenian draws between two kinds of self-mastery: one that is compatible with the CONFLICT model of virtue, and one that is compatible with the HARMONY model of virtue. This distinction allows for a more straightforward reading of the text in that it provides the resources to show how the Athenian can consistently (and genuinely) endorse both the CONFLICT and HARMONY models of virtue. By effectively dissolving the dilemma about virtue that has occupied recent commentators, my interpretation faces none of their problems. It also brings to light latent philosophical resources that have not been previously noted or appreciated. In order to advance my interpretation, it will be necessary to return to the text of the *Laws*.

§3 SELF-MASTERY & MODERATION

In the third book of the *Laws*, the Athenian details the history of three Greek states: Argos, Messene and Sparta.[161] The upshot in doing so is that it will allow the Athenian to examine some legislation without needing to engage "in idle speculation, but [by] investigating what has actually and truly happened" (684a1). As it turns out, these three states were ruled by three *brothers* (685d4).[162] These brothers initially

[159] Meyer (2018), p. 108

[160] The Athenian also claims that "the tale gives us a more lucid articulation of virtue and vice" (645c1). But the payoff of this "greater clarity" is that it will shed light on education and other practices like drinking parties.

[161] Cf. *Laws* 692d3, where Sparta seems to represents *reason*, Argos *spirit*, and Messene *appetite*.

[162] These brothers are the offspring of a demigod. This directly recalls the earlier analysis from 628a1. These kings are not only said to possess common subjects, but also a common army, which forms a "single unified body" (685d3).

exchanged oaths in accordance with mutually binding laws so that their kingdoms were ultimately "brought under the control of a single family" (686a4). Unfortunately, their alliance quickly evaporated: two of the three brothers became filled with greed (691a4).[163] Only one of them (the Spartan) continued to honor the common oath which had originally bound and united all three of them. Now, if it were not already clear that this is meant to recall the familial case of conflict from the first book of the *Laws*, the Athenian loses any pretense to subtlety when he again reminds his interlocutors—not even twelve lines later—that these kings are brothers (686a4). Consistent with his earlier assessment, the Athenian now claims that it would have been best if these three brothers had remained in harmony with one another (693a7). Unfortunately, their "lack of harmony" (691a7) resulted in conflict, which eventually led to the destruction of their familial empire.

The point in recounting these events, the Athenian now informs his interlocutors, is that if they can determine the cause of this unfortunate dissolution, then they will be in a much better position to avoid similar pitfalls when enacting their own legislation. He here reminds them that their legislation should be constructed with a view to virtue (693b2). But he now adds that legislation constructed with a view to "moderation […] wisdom or friendship" (τὸ σωφρονεῖν […] φρόνησιν ἢ φιλίαν, 693c3) are all equally acceptable ways of stating this same aim or goal. Indeed, he is careful to note that "all these aims are the same, not different" (693c3). If we think they are different, he tells us, then we must also take pains to figure out in what respect they are the same (693c3). In the next two sub-sections, I explore the respects in which the legislative aim of moderation is both different from and the same as wisdom and virtue. My reason for doing so is simple: moderation turns out to be a kind of self-mastery.

3.1 Moderation: Necessary or Sufficient for Virtue?

In the ensuing discussion, the Athenian applauds Sparta for refusing to confer civic distinction or office on the basis of such superficialities as wealth, good looks, or even the possession some particular virtue (like courage) if that virtue is not also accompanied by *moderation* (σωφροσύνη) (696b5).[164] He explains that even though courage is "one part of virtue" (696b7) no one would want to have someone who is courageous but immoderate (ἀκόλαστον) living in their home or in their community.[165] Without moderation, it seems, no other virtue is much to speak of:

> But surely, in the absence of moderation, justice will never spring up [...] Nor will the wise person we just now mentioned, whose pleasures and pains (ἡδονὰς καὶ λύπας) are in harmony with right reason and follow it (συμφώνους τοῖς ὀρθοῖς λόγοις καὶ ἑπομένας)." (696c5)[166]

---

[163] Greed (τὸ πλεονεκτεῖν) is a term that Plato tends to reserve for those ruled by their *appetitive* desires.

[164] Cf. *Laws* 630b1; Sparta was the only state which continued to honor their original agreement with Argos and Messene.

[165] Moderation is one of Plato's four canonical virtues—along with wisdom, justice, and courage. These are what the Athenian also refers to as 'divine goods' (631b7). Unlike the other three virtues however, there is still no standardly accepted translation for the virtue I am calling moderation. This virtue is just as often rendered 'temperance,' or 'prudence' or even 'soundness of mind.' It is even sometimes rendered 'self-control' or 'self-mastery.' This is especially true in the Saunders and Schofield translations of the *Laws*. Cf. Meyer's review of Schofield (2018)

[166] Ἀλλὰ μὴν τό γε δίκαιον οὐ φύεται χωρὶς τοῦ σωφρονεῖν [...] Οὐδὲ μὴν ὅν γε σοφὸν ἡμεῖς νυνδὴ προυθέμεθα, τὸν τὰς ἡδονὰς καὶ λύπας κεκτημένον συμφώνους τοῖς ὀρθοῖς λόγοις καὶ ἑπομένας.

So moderation is at least *necessary* for virtue.[167] But is it *sufficient* for virtue? Or, as the Athenian himself puts it:

> If we found moderation (σωφροσύνη) existing in the soul without all of the other virtues (ἄνευ πάσης τῆς ἄλλης ἀρετῆς), should we be justified in admiring it, or not? (696d4)[168]

Megillus demurs—he doesn't know whether they would be justified in admiring it. But the Athenian approves of this non-answer: "a very proper reply, for if you had opted for either alternative (ὁποτερονοῦν) it would have struck an odd note, I think" (696d7). The reason it would have struck an odd note is because moderation on its own is not entirely unworthy of admiration, but neither is it all that valuable when separated from the preeminent virtue of wisdom. Without wisdom, moderation turns out to be a mere "adjunct that isn't worth saying much about" (696d11).

3.2 Two Kinds of Moderation

While moderation without wisdom isn't worth saying *much* about, it is still apparently worth saying a *little* about, which is what the Athenian goes on to do in the fourth book of the *Laws*. When describing the conditions under which a legislator could most effectively enact laws, the Athenian maintains that it would be in a state ruled by a tyrant (709e7). This tyrant would ideally be young, have a good memory, a quick wit, and a character of natural elevation.[169] Yet if these qualities are to be useful to the legislator, the soul of this young tyrant should also possess "that quality which in our earlier discussion we said must accompany all the parts of virtue" (710a2). This quality, of course, was moderation. But the Athenian now distinguishes between two different kinds of moderation:

> I mean the popular kind (τὴν δημώδη), Clinias, not the exalted kind (σεμνύνων) one might invoke when insisting that moderation is also wisdom (φρόνησιν προσαναγκάζων εἶναι τὸ σωφρονεῖν). I have in mind the quality that naturally develops in children and animals—some of whom are akratically disposed with respect to pleasures (τοῖς μὲν ἀκρατῶς ἔχειν πρὸς τὰς ἡδονάς), others enkratically (τοῖς δὲ ἐγκρατῶς). We said that if this quality existed in isolation from the many other

[167] Justice will never spring up without moderation because justice was defined as the conjunct of wisdom, moderation and courage (631c5): "wisdom itself is the leading 'divine' good; second comes the habitual moderation of a soul that uses reason. If you combine these two with courage, you get (thirdly) justice; courage itself lies in fourth place."

[168] Σωφροσύνη ἄνευ πάσης τῆς ἄλλης ἀρετῆς ἐν ψυχῇ τινι μεμονωμένη τίμιον ἢ ἄτιμον γίγνοιτ' ἂν κατὰ δίκην;

[169] These are the qualities that pick out a philosophic nature in the *Republic* (474b2).

so-called goods we are discussing, it was not worthy of mention. (710a5)[170]

It is clear that the popular kind of moderation is the quality that the Athenian had earlier referred to as a mere 'adjunct.' This kind of moderation is now said to develop naturally in children and animals. It is described as a kind of self-control or restraint with respect to pleasure. Some children and some animals are naturally more *enkratic*—naturally better able to control or restrain the pull of pleasure—than others.[171] On could think here of children faced with the prospect of unlimited candy, or puppies faced with the prospect of unlimited kibble: some children and some puppies will eat themselves sick, while others will not.[172]

Indeed, even in the first book of the *Laws* (before the puppets passage), the Athenian had seemed to allow for this popular kind of moderation. He had there invoked the "moderate disposition of soul" (631c7) that needs to be combined with wisdom in order to be counted among the 'divine' goods, or virtues. When discussing this moderate disposition in separation from wisdom, the Athenian consistently described it as a natural tendency to restraint in the face of pleasure (634a–b, 635e–636e, 636c7). Indeed, the Athenian also suggests that moderation is perfected only after repeatedly practicing such restraint:

> So will anyone become perfect (τελέως) in moderation if they haven't done battle against the many pleasures and desires that urge them to commit shameless and unjust actions—if they haven't defeated (νενικηκὼς) them by dint of reason (μετὰ λόγου), effort, and skill, both in play and in earnest? (647d3)[173]

The language here is that of battle, and *victory* in this battle is construed as effectively resisting pleasure. This, of course, is exactly how self-mastery is described.

Now, recall that the puppets passage was a tale of virtue, which promised to help make the meaning of self-mastery clearer (645b2). The kind of self-mastery it seemed to model was the sort wherein reason emerges victorious over the opposing pulls of pleasure and pain. Here, a popular kind of moderation turns

---

[170] καὶ οὐχ ἥν τις σεμνύνων ἂν λέγοι, φρόνησιν προσαναγκάζων εἶναι τὸ σωφρονεῖν, ἀλλ' ὅπερ εὐθὺς παισὶν καὶ θηρίοις, τοῖς μὲν ἀκρατῶς ἔχειν πρὸς τὰς ἡδονάς, σύμφυτον ἐπανθεῖ, τοῖς δὲ ἐγκρατῶς· ὃ καὶ μονούμενον ἔφαμεν τῶν πολλῶν ἀγαθῶν λεγομένων οὐκ ἄξιον εἶναι λόγου.

[171] When discussing what it means to be 'self-defeated' in the first book of the *Laws*, Clinias claims that "we are all much more likely to mean someone defeated by pleasure than by pains" (633e4). This suggests that the popular meaning of self-mastery will likely be more concerned with pleasure than pain, which is exactly what we find. (Cf. *Republic* 389d7, 430e6)

[172] One might here object that children (and *a fortiori*, animals) do not possess reason—at least not in any robust sense. It would thus be strange to describe the sort of moderation they possess as reason 'controlling' or 'mastering' appetitive desires. It is worth noting, however, that the Athenian is prepared to call animals and children naturally *enkratic* (710a5). Moreover, it is clear that the Athenian thinks that this natural sort of control or restraint—at least in the case of children—eventually can (and should) be informed by reason (cf. 963e3 for a parallel discussion about courage). So perhaps it is better to think of the sort of control over pleasure that some children and animals exhibit as a kind of proto-control or proto-victory over pleasure. This kind of natural control is the sort of thing *capable* of being directed by reason in humans. Actualizing this capacity is arguably the role of education and habituation. Cf. 645e ff. for a discussion of drinking, which returns us to the state of young children. In the state of drunkenness, our cognitive abilities abandon us and we are then least in control of ourselves.

[173] Σώφρων δὲ ἄρα τελέως ἔσται μὴ πολλαῖς ἡδοναῖς καὶ ἐπιθυμίαις προτρεπούσαις ἀναισχυντεῖν καὶ ἀδικεῖν διαμεμαχημένος καὶ νενικηκὼς μετὰ λόγου καὶ ἔργου καὶ τέχνης ἔν τε παιδιαῖς καὶ ἐν σπουδαῖς

out to model a similar kind of self-mastery—the kind that is exhibited by those who are naturally disposed to achieve a kind of *victory* over pleasure. For this reason, the popular kind of moderation seems to conform nicely to the CONFLICT model of virtue.

But what of perfect moderation—the exalted kind that is also wisdom?[174] Does it, too, model a sort of self-mastery? It is clear that the Athenian is appealing to the connection between 'σωφρονεῖν' and 'φρόνησις.' This connection is evident both in their common etymology, and in their ordinary usage, where 'σωφρονεῖν' (the verb cognate with 'σωφροσύνη') regularly means "to be wise."[175] It is also clear that this exalted kind of moderation is what the Athenian had earlier paired with "wisdom or friendship" (φρόνησιν ἢ φιλίαν)—when all three were said to capture the same legislative aim as virtue (693c3). But we should also recall the Athenian's repeated insistence that the pleasures and pains of the wise person are in *harmony* with right reason and follow it (653b2, 696c5).[176] The person who possesses the exalted kind of moderation is wise, and so has achieved harmony or friendship between reason and feeling of pleasure and pain.[177] For this reason, the exalted kind of moderation seems to conform nicely to the HARMONY model of virtue. My suggestion is that exalted moderation also models a kind of self-mastery—the kind of self-mastery exhibited by those who attain *harmony* between reason and feelings of pleasure and pain.

§4 TWO KINDS OF VIRTUE

So we are now armed with an account of two kinds of moderation: a popular kind that seems to conform to the CONFLICT model of virtue, and an exalted kind that seems to conform to the HARMONY model of virtue. In order to bolster this account, I now want to return to the first book of the *Laws*, where moderation was originally introduced. Moderation places second, after wisdom, among the four divine goods or virtues (631c6).[178] Once the survey of these virtues is complete, the Athenian informs his interlocutors that—after enacting laws that seek to promote these virtues—the legislator will appoint guardians charged with protecting the laws. It turns out that these guardians come in two varieties:

> Some of whom possess wisdom (τοὺς μὲν διὰ φρονήσεως), others of whom possess true opinion (τοὺς δὲ δι' ἀληθοῦς δόξης ἰόντας), so that intellect (ὁ νοῦς) will bind everything together to follow moderation and justice (σωφροσύνῃ καὶ δικαιοσύνῃ) rather than wealth and ambition." (632c5)[179]

Two sorts of guardians will be appointed: some will possess wisdom, and some will not.[180] Now, we saw

---

[174] The account of moderation as a divine good at *Laws* 631c7 also fits this picture.

[175] Hence "know thyself" is a dictum of σωφροσύνη.

[176] See pp. 6, 11 above for discussion of these two passages.

[177] Cf. *Laws* 689d4: "Without harmony (ἄνευ συμφωνίας), how could there be wisdom (φρονήσεως) of even the smallest degree (τὸ σμικρότατον εἶδος)? There is no way (οὐκ ἔστιν)."

[178] Cf. *Laws* 696e3

[179] τοὺς μὲν διὰ φρονήσεως, τοὺς δὲ δι' ἀληθοῦς δόξης ἰόντας, ὅπως πάντα ταῦτα συνδήσας ὁ νοῦς ἑπόμενα σωφροσύνῃ καὶ δικαιοσύνῃ ἀποφήνῃ, ἀλλὰ μὴ πλούτῳ μηδὲ φιλοτιμίᾳ.

[180] Cf. *Republic* 506c7: "Haven't you noticed that opinions without knowledge are shameful and ugly things? The best of them are blind—or do you think that those who express a true opinion without understanding are any different from blind people who happen to travel the right road?"

in the previous section that the kind of moderation that is separated from wisdom is barely worth mentioning—let alone praising (696d11, 710a5). So at this juncture it is worth asking the following question: what sort of moderation will be possessed by those guardians who lack wisdom?

Before answering this question, it is worth briefly pointing out some other dialogues wherein the virtue of moderation is discussed in separation from wisdom. In the *Phaedo*, for instance, the sort of moderation exhibited by those without wisdom is referred to as mere "popular and political virtue (τὴν δημοτικὴν καὶ πολιτικὴν ἀρετὴν)" (82b1)—the sort of virtue "instilled by habituation and practice without philosophy and without intellect (ἐξ ἔθους τε καὶ μελέτης γεγονυῖαν ἄνευ φιλοσοφίας τε καὶ νοῦ)" (82b2). In the *Republic*, too, the sort of moderation possessed by non-philosophers is counted among the "popular virtues" (τῆς δημοτικῆς ἀρετῆς)" (500d8).[181] A bit later on, Socrates claims that the person who forcibly holds their appetites in check:

> Wouldn't be entirely free from internal civil war […] though generally their better desires are in control (κρατούσας) of the worse […] For this reason, they'd be more respectable than many, but the true virtue (ἀληθὴς ἀρετὴ) of a unified and harmonious (ὁμονοητικῆς) soul far escapes them. (554d8)[182]

I point out these passages because I now want to suggest that the two accounts of moderation (popular and exalted) not only track a distinction between two types of self-mastery—but also track a distinction between two types of *virtue*. If we apply this distinction to the two sorts of guardians above, it quickly becomes clear that those who possess wisdom will possess *exalted* virtue, while those guardians who do not will possess *popular* virtue.[183] But note that *all* of the guardians will possess some kind of 'virtue' on my account.[184] This, I take it, is a welcomed result. Surely those tasked with protecting the laws enacted by the legislator—laws that are ultimately constructed with a view to virtue (693b2)—should themselves be virtuous.

On the interpretation I am advancing, then, the puppets passage remains a 'tale of virtue.' It illuminates the *popular* kind of virtue, which issues from the popular kind of moderation or self-mastery.[185] This would explain why the Athenian explicitly *qualifies* what he takes the puppets passage to reveal about self-mastery: it only makes the meaning clear 'in a way' (τρόπον τινὰ).[186] It would also explain why the Athenian encourages his interlocutors to think of virtue as a sort of self-mastery: because (in a way) he thinks it is.

---

[181] Cf. *Statesman* 309c6 ff.

[182] Οὐκ ἄρ' ἂν εἴη ἀστασίαστος ὁ τοιοῦτος ἐν ἑαυτῷ […] ἐπιθυμίας δὲ ἐπιθυμιῶν ὡς τὸ πολὺ κρατούσας ἂν ἔχοι βελτίους χειρόνων […] Διὰ ταῦτα δὴ οἶμαι εὐσχημονέστερος ἂν πολλῶν ὁ τοιοῦτος εἴη· ὁμονοητικῆς δὲ καὶ ἡρμοσμένης τῆς ψυχῆς ἀληθὴς ἀρετὴ πόρρω ποι ἐκφεύγοι ἂν αὐτόν. Cf. *Republic* 430e9

[183] Cf. Vasiliou (2014), p. 21. The guardians who lack wisdom will be akin to what Vasiliou calls *Phd-philosophers*.

[184] *Contra* Meyer (2018), who argues that only one of the two models can ultimately be the correct account of virtue.

[185] Cf. *Meno* 98b1–c1, where knowledge and true opinion are equivalent in their practical effects on behavior. In the *Republic*, this tracks the distinction between good guardians and good watchdogs.

[186] Indeed, the treatment of self-mastery in the puppets passage coheres well with the way that self-mastery it is treated in other dialogues, most notably the *Republic*, where both moderation and self-mastery are initially glossed as a sort of control over pleasure (430e6). When analyzed more deeply, however, their meanings are *refined*—just as they are in the *Laws*. Cf. *Laws* 633e3. See also *Gorgias*, where Socrates defines moderation as the state in which a person "rules the pleasures and appetites within himself" (491d11–e1).

So while I accept the rather elegant suggestion of the third interpretive strategy that the CONFLICT model of virtue represents a developmental stage on the way to the HARMONY model of virtue, I reject the assumption common to all three interpretive strategies that these two models of virtue are incompatible with one another—and that we must therefore choose between them.[187] On my account, the CONFLICT model captures the popular kind of virtue and the HARMONY model captures the exalted kind of virtue. Indeed, this should not be surprising given that the puppets passage takes off from Clinias' initial thesis that self-mastery is a kind of victory (626e2). This is the popular self-mastery ultimately captured by the CONFLICT model of virtue and which, at best, renders someone enkratically disposed toward pleasure. It is not until much later that we receive an account of exalted virtue—and the exalted kind of moderation exhibited by those who possess it. The Athenian was thus not wrong when he suggested that an upshot of his 'tale of virtue' was that it would eventuate in a clearer understanding of virtue and vice (645c1). He was also not wrong to suggest that the meaning of 'self-mastery' would eventually become clearer. Exalted moderation is *perfected* self-mastery.

BIBLIOGRAPHY

Annas, Julia. "Virtue and Law in Plato" in *Plato's Laws: A Critical Guide* (2010: Cambridge University Press)

Barney, Rachel. *Plato and the Divided Self* (2014: Cambridge University Press)

Belfiore, Elizabeth. "Wine and the Catharsis of Emotions in Plato's Laws." *Classical Quarterly* (1986): 421–37

Bobonich, Christopher. "Akrasia and Agency in Plato's *Laws* and *Republic*" in *Essays on Plato's Psychology* (2001:
Lexington Books)

Bobonich, Christopher. *Plato's Laws: A Critical Guide* (2010: Cambridge University Press)

Bobonich, Christopher. "Plato on Akrasia & Knowing Your Own Mind" in *Akrasia in Greek Philosophy* (2007:
Brill)

Bobonich, Christopher. *Plato's Utopia Recast: His Later Ethics and Politics* (2002: Clarendon)

Brisson, Luc. "Ethics and Politics in Plato's *Laws*," *Oxford Studies in Ancient Philosophy* (2005): 93–121

Brisson, Luc, and Scolnicov, Samuel (eds.). *Plato's Laws: From Theory Into Practice: Proceedings of the VI Symposium*
Platonicum (2003: Academia Verlag)

[187] Meyer (2018), p. 108: "repeated success at resisting the pull of opposing desires and fears will ultimately result in retraining those desires and fears so that they agree with rather than oppose the golden cord."

Callard, Agnes. “Akratics as Hedonists,” *Ancient Philosophy* [forthcoming]

Cooper, John. *Plato: Complete Works* (1997: Hackett)

Fine, Gail. *The Oxford Handbook of Plato* (2008: Oxford University Press)

Frede, Dorothea. “Puppets on Strings: Moral Psychology in Laws Books I and II” in *Plato’s Laws: A Critical Guide* (2010: Cambridge University Press)

Gerson, Lloyd. “Akrasia and the Divided Soul in Plato's *Laws*” in *Plato's Laws: From Theory into Practice* (Proceedings of the VI Symposium Platonicum)

Gerson, Lloyd. *Knowing Persons: A Study in Plato* (2006: Clarendon Press)

Gerson, Lloyd. “Knowledge and the Self in Platonic Philosophy” in *Proceedings of the Boston Area Colloquium in Ancient Philosophy* (2000): 231–53

Gerson, Lloyd. “Plato’s Rational Souls,” *The Review of Metaphysics* (2014): 37–59

Gerson, Lloyd. “Plotinus on Akrasia: The Neoplatonic Synthesis,” in *Weakness of Will, From Plato to the Present* (2008: Catholic University of America Press)

Gerson, Lloyd. “Virtue With and Without Philosophy” in *festschrift for John Rist* (forthcoming)

Kahn, Charles. “From *Republic* to *Laws*: A Discussion of Christopher Bobonich, *Plato's Utopia Recast*,” *Oxford Studies of Ancient Philosophy* (2004): 337–362

Kahn, Charles. “On Platonic Chronology” in *New Perspectives on Plato, Modern and Ancient* (2002: Harvard University Press)

Kamtekar, Rachana. “Psychology and the Inculcation of Virtue in Plato’s Laws” in *Plato’s Laws: A Critical Guide* (2010: Cambridge University Press)

Klaus Schöpsdau, *Nomoi Buch VIII-XII (Gesetze) Platon Werke, Übersetzung und Kommentar* (2003: Göttingen)

Nightingale, Andrea. “Plato’s Law Code in Context: Rule by Written Law in Athens and Magnesia,” *Classical Quarterly* (1999): 100–22

Mackenzie, Mary Margeret. *Plato on Punishment* (1981: University of California Press)

Mayhew, Robert. “Persuasion and Compulsion in Plato’s *Laws*,” *Polis* (2007): 91–111

Meyer, Susan Sauvé. *Plato: Laws 1 & 2, translated with a commentary* (2015: Clarendon Press)

Meyer, Susan Sauvé. *Review: Akrasia in Greek Philosophy: from Socrates to Plotinus* (2008: Notre Dame Philosophical
Reviews)

Meyer, Susan Sauvé. *Review: The* Laws*, Cambridge Texts in the History of Political Thought* (2018: Bryn Mawr Classical
Review)

Meyer, Susan Sauvé. "Self-Mastery and Self-Rule in Plato's Laws" in *Virtue, Knowledge, and the Good: essays in*
honour of Gail Fine and Terence Irwin (2018: Oxford University Press)

Morrow, Glenn. *Plato's Cretan City* (1993: Princeton University Press)

O'Brien, Michael. "Plato and the 'Good Conscience': *Laws* 863e5–864b7, *Transactions and Proceedings of the*
American Philological Association (1957): 81–7

Riesbeck, David. Review of Meyer 2015. *Notre Dame Philosophical Reviews*. 24 May 2016.

Sassi, *Maria Michela.* "The Self, the Soul, and the Individual in the City of the Laws," *Oxford Studies in Ancient*
Philosophy (2008): 125–48

Saunders, Trevor. "The Structure of the Soul and the State in Plato's *Laws*," *Eranos* (1962): 37–55

Schofield, Michael. "Plato's Marionette," *Rhizomata* (2016): 128–53

Schofield, Michael. *Plato: The Laws* (2016: Cambridge University Press)

Stalley, R. F. *An Introduction to Plato's Laws* (1983: Hackett)

Stalley, R. F. "Justice in Plato's *Laws*" in *Plato's Laws: From Theory Into Practice: Proceedings of the VI Symposium*
Platonicum (2003: Academia Verlag)

Stalley, R. F. "Persuasion in Plato's *Laws*," *History of Political Thought* (1994): 157–77

Vasiliou, Iakovos. "From the *Phaedo* to the *Republic*: Plato's Tripartite Soul and the Possibility of Non-Philosophical Virtue" in *Plato and the Divided Self* (2014: Cambridge University Press)

Wilburn, Joshua. "*Akrasia* and Self–Rule in Plato's *Laws*," *Oxford Studies in Ancient Philosophy* (2012): 25–53

Wilburn, Joshua. "Moral Education and the Spirited Part of the Soul in Plato's *Laws*," *Oxford Studies in Ancient*
Philosophy (2013): 63–102

Moral Responsibility & Cooperation

Saba Bazargan-Forward

1. Background

We share an action when we coordinate and combine our individual actions in furtherance of doing something together, as when we walk together or paint a house together.[188] Shared actions can be comprised of many participants, each whom contributes very little to what they together do. Some have argued that in such cases, each participant is morally responsible not just for the difference she makes, but for the collective action *in toto*.[189]

Such a view, however, requires argument. We cannot just stipulate that participants in shared action are morally responsible for more than the difference she makes. To treat this as a bedrock claim is problematic since the intuition seems to conflict with the equally plausible pronouncement that each of us is morally responsible only for what is within her own causal reach. What we need is an account explaining how and why cooperation in the context of shared action dramatically expands the scope of moral responsibility. I provide such an argument here.

Typically, participants in shared action take themselves to have claims against one another that they 'do their part'; I will argue that this mutually presumed claim inculpates each participant in what every other participant does *qua* participant. As a result, each individual can, in principle, be morally responsible for what they do together, irrespective of the degree to which they contribute to the shared action.

If you and I are undertaking a shared action, the presumed claim that you have against me that I do my part, and the presumed claim that I have against you that you do your part, derives from the mutual agreements (explicit or implicit) we make at the onset of our shared activity. By so agreeing, each participant enjoys a limited kind of presumed authority over the other. The authority is *limited* in that the shared action needn't necessarily involve mutual micromanagement: each participant has a claim against the other that they do their part, without necessarily specifying what that consists in. And the authority is *presumed* in that if the shared action is morally wrongful then each participant will at best *incorrectly believe* that she has a claim against others that they do their part. If you and I agree to rob a bank together, that agreement obviously does not give me a claim against you that you do your part in the bank robbery. But if, as a result of our agreement to rob the bank, each of us *believes* that we have such a claim against each other, then we enjoy *presumed* authority over one another – and that suffices as a basis for mutual inculpation – or so I will argue. (For the sake of expository convenience, I will suppress the caveat "presumed" from here on).

In the context of shared action, the authority we have over another yields a *protected reason* to act accordingly. A protected reason to do ϕ is comprised of a first-order reasons to do ϕ and a second-order reason to exclude certain competing considerations from deliberation pertaining to ϕ.[190] So, if you and I are engaged in shared action, I will typically have authority over you that you do your part and you will typically have authority over me, that I do my part, where this authority is cashed out in terms of a protected reason. In arguing for this view, I will focus on the two most influential accounts of shared action: Margaret Gilbert's and Michael Bratman's.

[188] To take canonical examples from (Gilbert, 1990) and (Bratman, 1992).
[189] See (Lepora & Goodin, 2013, p. 8), (Miller, 2006, pp. 177-78, 181), (Haque, 2017, pp. 57, 59, 263), among others who adopt versions of this view.
[190] See (Raz J. , 1977), (Raz J. , 1990, pp. 35-84).

I then argue that where I have authority over you that you do ϕ, and where you subsequently do ϕ, thereby conforming to the protected reasons you take there to be, I thereby end up morally responsible for what you do. Of course, if you choose to do ϕ in part because of me, then I might be morally responsible for what you do in virtue of my causal influence on your voluntary conduct (intervening agency notwithstanding). But I argue that there is another, separate basis for thinking that I am morally responsible for what you do. Irrespective of my causal contributions to your conduct, the very fact that I have presumed authority over you that you do ϕ can ground my moral responsibility for what you do should you abide by that claim I have against you.

Why is this so? At its most basic level, rational agency can be divided broadly into *deliberative* and *executory* functions, where the former is the process by which options are evaluated and selected, and the latter the process by which the selected options are implemented. When an agent makes a practical decision, she transitions from the deliberative process to the executory process, in that the function of a decision is to enact via conduct the practical reasons the agent takes there to be.

Normally, a single agent embodies both the deliberative and executory functions of agency. That is, normally, *you* deliberate by evaluating and selecting among candidate options, and *you* consequently enact the option that you have selected. It is possible, though, for you to 'outsource' the executory functions to *me*. In such a case, you attribute to me the role of enacting the practical reasons you take there to be; should I accept that role, you thereby change the object at which your practical reasoning is teleologically directed. Its object is no longer *your* conduct, but rather *mine*. That is to say, the practical reasons *you* take there to be have the function of guiding *my* conduct, and *my* conduct has the function of enacting the practical reasons *you* take there to be. In this way, we establish an interpersonal division of agential labor, in that you count as the 'decider', and I count as the 'doer'. The purpose of establishing an interpersonal division of agential labor, then, is to separate out and assign to multiple agents the deliberative and executory functions of rational agency, so that it is the role of one agent to evaluate and select among candidate options, and the role of another agent to implement that option via conduct.

I argue that we establish a division of agential labor of this sort, when you grant me, and I accept, authority over you pertaining to some act ϕ, as a result of which I have a claim against you that you do ϕ, where that claim yields a protected reason for you to do ϕ. In virtue of the reason's protected status, the practical reasons I take there to be has the function of guiding your conduct, and your conduct has the function of enacting those practical reasons.

We morally assess conduct partly by assessing the practical reasons for which it was done. So, to morally assess *your* conduct, we need to repair to the practical reasons *I* take there to be – given the functional relationship between your reasons and my conduct. If my practical reasons are morally problematic, then I am responsible for a wrong-making feature of your conduct. The result, then, is that I am at least partly morally responsible for what you do (though my responsibility in no way diminishes yours).

To be clear, the claim here is not that I am morally responsible for what you do in virtue of having influenced your motivations (though that too might be a basis for moral responsibility). Rather, the basis of my responsibility for your conduct is constitutive rather than causal. *Your conduct has the function of enacting the practical reasons I take there to be; so, by adopting morally problematic practical reasons, I constitutively determine, from afar, the purpose of your conduct.* Insofar as that affects the moral assessment of your conduct, I am on the hook for that difference I make. I call this 'authority-based moral responsibility'.

The upshot is this. Any given participant in typical instances of shared action has authority over the other participants that they do their part, in that she has a claim against them yielding a protected reason for them to do their part. The practical reasons she takes there to be has the function of normatively guiding

their conduct in that regard, and their conduct has the function of enacting the practical reasons she takes there to be. So, given the argument for authority-based responsibility, she is morally responsible for what they do should they conform to those reasons. Because this is true of every participant, the result is that each participant can be morally responsible for what they together do – irrespective of the degree to which they causally contribute to what they together do.

2. Presumed Authority and Shared Action

Normally, you decide whether to do some action ϕ by considering the reasons for and against ϕ. Things are different, though, when you decide what to do in response to someone you believe to be authorized to tell you what to do with respect to ϕ. The reason to do ϕ in such a case derives not from the merits of ϕ itself, but from the very fact that the authority has a claim against you that you do ϕ.[191] In addition, the presumed authority's claim against you that you do ϕ provides a reason to exclude certain competing first-order reasons from further deliberation pertaining to ϕ. (The authoritative claim does not exclude *all* competing reasons, however; the zone of exclusion varies with the command and the context in which it is given.)[192] The combination of the first-order reason to act in accordance with the authoritative claim against you, and second-order reason to exclude from deliberation certain competing first-order reasons, yields what Joseph Raz calls a *protected* reason.[193]

Deference to authority, then, requires doing what the authority commands precisely because the authority commands it. In this respect, the commands of an authority serves to 'settle the matter' for you within a particular domain of conduct.

There is, of course, a vast literature on what constitutes legitimate authority, and even whether such authority is possible, especially in the context of whether governments are justified in compelling its citizen to obey laws. But here I limit myself to authority between individuals. For example, one way for an individual to gain authority over you is by accepting a promise you made; the reasons to fulfill a promise are in general protected.[194] So, if you promise to me that you will do ϕ, then I have authority over you with respect to ϕ, by deciding whether or not to hold you to that promise. If I decide to hold you to that promise, then I have a claim against you that you do ϕ, where that claim provides you with a protected reason to do ϕ.[195] (Of course, if the promise is immoral, then you at best might only *believe* that you have a protected reason to enact it; in these cases, the authority is merely presumed. I continue to suppress that caveat in what follows).

I will argue that participants in shared action have authority over one another – even in cases of shared action *among equals*. The participants have authority over one another in that each participant has a claim against every other participant that they do their part in the shared action, where that claim yields a protected reason to comply.

3. Authority-Based Moral Responsibility

In what follows I will argue that when we both believe that I have an authoritative claim against you that you do ϕ, and you subsequently conform to that claim, I am morally responsible for what you do irrespective of my causal influence on your conduct (which isn't to say that you are any less responsible for what you do).

3.1. Establishing an Agential Division of Labor

[191] (Hart, 1990, p. 101) For helpful discussion see (Shapiro, 2002), (Owens, 2008), and (Westlund, 2011).
[192] For helpful discussion, see (Shapiro, 2002, pp. 406-7) and (Owens, 2008).
[193] See (Raz J. , 1977), (Raz J. , 1990, pp. 35-84).
[194] See (Raz J. , 1986, pp. 35-7).
[195] See Owens (2008), (2012) for an extended discussion of promise-making as conferring authority.

Suppose a mafioso asks a hitman to assault a particular innocent. The hitman freely promises to do so. In this case, the mafioso and the hitman have established an agential division of labor, in that the mafioso counts as the 'decider' and the hitman counts as the 'doer'. The mafioso counts as the decider in the sense that he has the function of normatively determining whether the hitman is to commit the assault, by either holding him to his promise or releasing him from it. And the hitman counts as the doer in the sense that the hitman's function is to implement the mafioso's decision pertaining to assault.

But what does it mean for someone to have a 'function'? And how does someone attain a function? There are, at the broadest level, two kinds of functions – natural and agentive. We *create* agentive functions. As John Searle puts it, these are "functions that we do not discover, and that do not occur naturally, but that are assigned relative to the practical interests of conscious agents".[196] For example, a chair, a car, and a wrench have the functions that they do in virtue of our intentions – specifically, our intentions to put these objects to use in furtherance of achieving particular tasks.

An agentive-function is *design-based* if the object possessing that function was created with the intention to use it in furtherance of some purpose. Alternatively, an agentive-function is *use-based* given an actual or dispositional use of that object in furtherance of some purpose. I will focus here on use-based functions.[197]

It's possible for *your* decisions to serve as a use-based agentive function for *me*. For example, suppose that in certain situations I flip a coin to determine which of two equally choiceworthy options to select. I thereby assign to that coin-flip a one-off, use-based, agentive function of determining for me which of the two options to select. Lacking a coin, however, suppose I decide to use your decisions as my coin. Unbeknownst to you, I decide what to do based on whether you use a contraction in your next uttered sentence. In this sort of case, your agency plays a wholly *passive role* in fulfilling the agentive function I've assigned to you in that the agentive function plays no normative role in your deliberations.

In some cases, though, you might agree to adopt the agentive function by agreeing to act in accordance with the conduct the function specifies. In this case, your agency plays an *active* role in fulfilling the agentive function assigned to you, in that the agentive function does indeed play a normative role in your deliberations. That is, the function enjoins you to exercise your deliberative capacities in furtherance of fulfilling that function. I will accordingly call this a '*deliberative*, use-based, agentive function'. My use of the word 'function', *sans phrase*, should be understood as shorthand for 'deliberative, use-based, agentive function'.

All functions have an end – an aim at which the function is normatively directed. We can characterize a function in terms of a protected reason to act in accordance with the function's end. This elegantly captures the nature of a function: it imposes upon us a particular end and (possibly) a particular means by which to achieve that end, while simultaneously delimiting deliberative freedom by excluding some (and possibly all) competing considerations from the balance of reasons.

So, to say that the mafioso has the function of deciding whether the hitman is to commit the assault and that the hitman has the function of acting accordingly, is to say that they both treat the mafioso's directives to the hitman pertaining to whether he should commit the assault as *settling that matter* for the hitman, in that the mafioso's directives yield protected reasons for the hitman.

Recall, though, that protected reasons also characterize *authority*. Person *A* and person *B* believe that *A* has authority over *B* that *B* do ϕ if they both believe that *A*'s claim that *B* do ϕ provides a protected

[196] (Searle, 1997, p. 20)
[197] For much more detail on agentive functions see (Ludwig, 2018, pp. 140-141) to whom I am indebted here.

reason for *B* to do ϕ. So, one way to characterize the functional relationship between the mafioso and the hitman is in terms of the authority the mafioso has over the hitman. This is because both authority and functional roles are cached out in terms of protected reasons. The upshot, then, is this: the authority the mafioso has over the hitman, and the function each of them has with respect to each other, are both is grounded in the protected reasons resulting from the promise which the hitman makes and the mafioso accepts.

The idea that the mafioso's claim against the hitman settles the matter for the hitman – in that the hitman takes himself to have limited deliberative say in the matter – might seem to overstate the role that the promise plays in the hitman's motivational economy. Suppose that if fulfilling the promise were not beneficial to the hitman, he would renege on it. The hitman seems to be violating the deliberative injunctions imposed by his protected reasons, even if he ultimately decides to go through with the assault.

Nonetheless, so long as the hitman believes that the promise provides a protected reason to commit the assault, then by the hitman's own lights, the mafioso has the authority to settle the matter for him, and to exclude further deliberation about the matter. If he denies *that*, then he has not made a sincere promise. So long as they both believe that the mafioso has a claim against the hitman where that claim yields a protected reason for the hitman to commit the assault, then the functional relationships characterizing the division of agential labor between them remains in place, even if the hitman's commitment to the protected status of that reason is less than perfect.

In summary: the mafioso and the hitman establish an agential division of labor in which the mafioso functions as the decider and the hitman functions as the doer. Each has this function in virtue of their belief that the mafioso's claim against the hitman yields protected reasons for the hitman to act accordingly. That is, each has the function that they do in virtue of the presumed authority that the mafioso has over the hitman. And that authority is grounded in the promise that the hitman makes and the mafioso accepts.

Clearly, both the hitman and the mafioso are severally and fully morally responsible for the ensuing assault. The mafioso is responsible partly in virtue of the causal role he plays in convincing or motivating the mafioso to commit the assault. But in what follows, I will argue that the mafioso is also morally responsible on different grounds: in virtue of the role he plays as the 'decider' and the role the hitman plays as the 'doer'. This will have consequences for responsibility in the context of shared action in general.

3.2. Responsibility in an Agential Division of Labor

When the hitman commits the assault, his victim is entitled to an explanation of what happened. She has an interest in ascertaining the wrongdoer's *practical reasons* for acting in the way that he did. Whether and why the wrongdoer believed the harm to be warranted is morally relevant to assessing the wrongful conduct.

But what type of practical reason are morally relevant to assessing the conduct? To morally evaluate the hitman's conduct, his victim would need to know his belief-relative practical reasons for harming her. That is, she is entitled to demand that he explain what he thought spoke in favor of harming her. The relevant belief-relative practical reasons are those that the wrongful conduct in question had the function of enacting. The purpose of practical reasoning, is, after all, to normatively guide conduct by determining what ends to pursue and how to pursue them. Concomitantly, the purpose of conduct (or at least intentional conduct) is to enact the practical reasons the actor take there to be.

Clearly, the hitman's conduct had the function of enacting some of his own practical reasons of which there might be a variety. The hitman might have committed the assault partly because he promised to do

so, and partly he enjoyed it, and partly because he was paid to do so, and partly because he wanted to earn the mafioso's respect, and so on. Suppose, though, that like most people the hitman thinks that when we make a promise, we make a commitment to the person to whom we make the promise, and that this gives us a protected reason to fulfill the promise irrespective of whether we benefit from doing so. This isn't to say that the hitman is motivated solely or even mostly by that sense of commitment. Neither does it suggest that the hitman is so committed that he would follow through with the promise even if he thought doing so wouldn't benefit him. Rather, the hitman registers the commitment he made as *a* reason – albeit a defeasible one – to fulfill the promise.

So at least one of the reasons the hitman takes there to be in favor of committing the assault does not advert to the various benefits he receives from doing so. Rather, the mafioso's authoritative claim against the hitman that he commit the assault, itself provides a reason to do so, in virtue of the authority the hitman granted the mafioso by promising to commit the assault. (Of course, the reason is merely presumed, in that it is a reason the mafioso and hitman take there to be – I continue to suppress this caveat).

But if the relevant belief-relative practical reasons are those that the wrongful conduct in question had the function of enacting, then they will include not only the hitman's reasons, but the mafioso's as well, since the mafioso has the function of specifying whether the hitman is to commit the assault, and the hitman has the function of acting accordingly, given the division of agential labor the two of them established. Recall that these dovetailing functions arise from the promise which the hitman made and which the mafioso accepted; by accepting the promise, the mafioso thereby issues a protected reason for the hitman to commit the assault unless the mafioso demands otherwise. Recall also that the purpose of intentional conduct in general is to enact the practical reasons we take there to be. So, if the mafioso gives the hitman the go-ahead (or otherwise refrains from instructing the hitman to stand down), and the hitman subsequently commits the assault, the hitman's conduct will have had the function of enacting the practical reasons the mafioso takes there to be in favor of committing the assault.

This means that, if we evaluate conduct by addressing the belief-relative practical reasons that the conduct has the function of enacting, then we ought to evaluate the hitman's conduct *by adverting to the mafioso's practical reasons*. So, if the victim wants to know all the belief-relative practical reasons behind the decision to commit the assault, she needs to address not just the hitman, but the mafioso as well. In this way, the practical reasons the mafioso takes there to be serves as a basis by which we evaluate what the hitman does. To the extent that the practical reasons the mafioso takes there to be are morally problematic, *the mafioso is responsible for a wrong-making feature of the hitman's conduct.* After all, if we evaluate conduct by addressing the belief-relative practical reasons that the conduct has the function of enacting, and if the mafioso's belief-relative practical reasons are problematic, then the mafioso has made it so that the hitman's conduct is morally problematic as well.

To be clear, the claim here is not that the mafioso is morally responsible in virtue of having influenced the hitman's motivations (though that too might be a basis for moral responsibility). Rather, the basis of mafioso's responsibility for the hitman's conduct is *constitutive* rather than causal. The hitman's conduct has the function of enacting the practical reasons the mafioso takes there to be; so, by adopting morally problematic practical reasons, the mafioso, from afar, constitutively determines the purpose of the hitman's conduct. Insofar as that morally affects an assessment of the hitman's conduct, the mafioso is on the hook for that difference he makes.

If the mafioso is responsible for a wrong-making property of the conduct in this way, then the mafioso (in addition to the hitman) is a proper object of the victim's reactive attitudes in the response to the wrong she has suffered. The result is that when the hitman fulfills an authoritative claim the mafioso has against

him, the mafioso can be responsible for the harm the hitman inflicts quite apart from the causal role the mafioso plays in that harm. I will call this the argument for 'authority-based moral responsibility'.

Of course, in the example under consideration, the mafioso is also *a cause* of the hitman's conduct; so it goes without saying that the hitman's victim can blame not just the hitman but the mafioso for the harm she has suffered. In this case, the authority-based grounds for moral responsibility is largely otiose. But in cases of shared action where each participant's causal contribution is morally insignificant, the authority-based grounds for moral responsibility will play an important role in inculpating the participants.

If this account is correct, then causal over-determination (whether concurrent or preemptive) in the context of shared activity is no impediment to responsibility. Suppose many individuals together agree to assist one another in furtherance of some shared action ϕ which each of them hopes to promote. Though each participant's contribution to ϕ is negligible, together they achieve ϕ, which, we can suppose, constitutes a substantial harm. On a standard, causal account of moral responsibility, any given participant is responsible at most for her contribution to what she causes, which by hypothesis is negligible. But given the argument for authority-based moral responsibility, any given participant will be potentially responsible for up to the sum total of what all the participants together do, provided that they are participants in shared action.

I've argued that the 'decider' in a division of agential labor is morally responsible for what the 'doer' when the latter follows the former's instructions. The basis for this responsibility is the authority the decider has over the doer, where the authority generates what each takes to be a protected reason for the doer to do as the decider instructions. The situation is fundamentally the same in cases of shared action among equals – that is, individuals in a non-hierarchical relationship. There is a sense in which they authority over one another.

It might seem strange to think that equals can have authority over one another. But in shared activity, the claim that each of us has over the other is that the other do his or her part, simply. That is, each member of the group has authority over every other member that they do their part in furtherance of the shared activity, whatever that part might be (either necessarily, as in Gilbert's account, or contingently as in Bratman's account). In this sense, when we undertake a shared action, each of us is a decider and a doer with respect to each other. I decide that the others will continue to do their part, and they decide that I will continue do my part; and I accordingly do what they decide, and they do what I decide.

So, if the argument for authority-based responsibility goes through in general, then it provides a basis for thinking that individuals acting together in the context of shared action are either responsible for what they do together, quite independent of their causal influence over each other's actions.

This isn't to say, however, that all participants in all instances of shared activity will always be fully inculpated. This is because protected reasons are scalar – the weaker the protection, the weaker the degree of inculpation. Recall that a protected reason to do ϕ is comprised of both a first-order reason to do ϕ, and a second order reason to exclude from deliberation certain reasons against doing ϕ. The constitutive first-order reasons obviously admit of degrees; but the exclusionary reasons do so as well. They admit of degrees by way of the size of the exclusionary zone the second-order reasons sets. Depending on how we specify the scalarity of the reasons the authoritative claims yield, there is room for the view that a participant bears little responsibility for what the vast majority of the other participants do – and yet is responsible for more than the difference she makes.

There are a host of remaining questions. Do *insincere* promises establish an inculpatory division of agential labor? What about acceding to mere requests? For want of space, I leave these issues for another time.

2019

Works Cited

Bratman, M. (1992). Shared Cooperative Activity. *The Philosophical Review, 101*(2), 327–341.

Gilbert, M. (1990). Walking Together: A Paradigmatic Social Phenomenon. *Midwest Studies in Philosophy, 15*(1), 1-14.

Haque, A. (2017). *Law and Morality at War.* Oxford: Oxford University Press.

Hart, H. (1990). Commands and Authoritative Legal Reasons. In J. Raz, *Authority* (pp. 92-114). New York City: New York University.

Lepora, C., & Goodin, R. E. (2013). *On Complicity and Compromise.* USA: Oxford University Press.

Ludwig, K. (2018). *From Plural to Institutional Agency* (Vol. II). Oxford: Oxford University Press.

Miller, S. (2006). Collective Moral Responsibility: An Individualist Account. In P. A. French (Ed.), *Midwest Studies in Philosophy* (Vol. XXX, pp. 176-193). Minneapolis: Wiley Blackwell.

Owens, D. (2008). Rationalism about Obligation. *European Journal of Philosophy, 16*(3), 403-431.

Owens, D. (2012). *Shaping the Normative Landscape.* Oxford: Oxford University Press.

Raz, J. (1977). Promises and Obligations. In P. Hacker, & J. Raz, *Law, Morality and Society: Essays in Honour of H.L.A. Hart.* Oxford: Clarendon Press.

Raz, J. (1986). *The Morality of Freedom.* Oxford: Oxford University Press.

Raz, J. (1990). *Practical Reasons and Norms.* Princeton: Princeton University Press.

Searle, J. (1997). *The Construction of Social Reality.* Free Press.

Shapiro, S. (2002). Authority. In J. Coleman, & S. Shapiro, *The Oxford Handbook of Jurisprudence* (pp. 382-339). Oxford: Oxford University Press.

Westlund, A. (2011). Autonomy, Authority, and Answerability. *Jurisprudence, 2*(1), 161-179.

What, if Anything, is Disagreement in Attitude?

Sarah Stroud

draft for NUSTEP—not for quotation or citation

Read one way, this question is quickly and easily answered. Belief is an attitude; so if there is such a thing as (what C. L. Stevenson called) disagreement in belief, then that is a kind of disagreement in attitude.

This easy question is not the one I will be exploring here, however. Stevenson introduced the idea of *disagreement in attitude* as a label for a phenomenon which is supposed to play an important role in (what would be, if successful) a powerful argument for non-cognitivism. We can thus view "disagreement in attitude" as a kind of placeholder or technical term referring to that thing, if any, which plays that particular functional role in the aforementioned argument. I will argue that once we understand the job that disagreement in attitude would need to do, it is unclear what, if anything, can do it. I suspect in short that the equation which functionally defines it lacks a solution.

I will start, as a paper on this topic should, with Stevenson, explaining the argument which I think is at work in his "The Nature of Ethical Disagreement" and spelling out what disagreement in attitude would have to be in order to accomplish the aim which it is introduced to serve. In the second section I characterize disagreement in general, building on its profile in the most familiar type of case. I then examine whether there is a species or subcategory of genuine disagreement with which we could identify disagreement in attitude. I suggest that there is indeed a phenomenon which responds to *some* of the desiderata for disagreement in attitude, functionally understood in terms of its role in the argument with which we began. But as I argue in the fourth section, the relevant phenomenon—which I call "practical disagreement"—is considerably narrower than the capacious phrase "disagreement in attitude" would suggest. More significantly, practical disagreement seems not to be up to the task of carrying all the burdens which disagreement in attitude would need to shoulder in order to serve Stevenson's argument in the way he intends. The job description for disagreement in attitude may remain unfilled.

I end by simply flagging a question which merits further exploration: whether any of the refinements or bells and whistles which were later added to the non-cognitivist conception of moral thought, opinion, and judgment are likely to change the basic conclusions which I drew in the paper. I don't think they will, although I don't pretend to justify that prognostication here.

1. A Powerful Argument for Non-Cognitivism

What is it to think a moral thought, to hold a moral opinion, to make a moral judgment?[198] The non-cognitivist, as I shall understand her here, offers a distinctive kind of answer to these questions in moral psychology. She has a view, that is, about what kind of mental or psychological phenomena the above are, namely that they are *not* cognitive, or doxastic, phenomena; rather, they are *conative*

[198] Since thoughts, opinions, and judgments are often assumed to be cognitive or doxastic phenomena by definition, it might be safer to ask: what is it to "think" a moral "thought," to "hold" a moral "opinion," to "make" a moral "judgment"? All these scare quotes, though, seem too fussy to continue: please take it as read that in asking these questions we are in fact remaining neutral about whether what we *call* moral "thoughts," "opinions," and "judgments" are genuinely doxastic phenomena.

phenomena.[199] I stress that I am interested here *only* in non-cognitivist answers to questions in moral psychology and will not be concerned with associated non-cognitivist views about moral language, moral terms, moral sentences, or other linguistic phenomena. Moreover, while I have put the point in terms of *moral* thoughts, opinions, and judgments, one might instead wish to offer a non-cognitivist account of *ethical* opinions (&c.), *normative* opinions (&c.), *evaluative* opinions (&c.), or "'ought' thoughts." I don't think it will matter to my discussion which of these is the non-cognitivist's target, so I will simply continue to speak of *moral* thoughts, opinions, and judgments as if that were the intended scope of her non-cognitivist claim (sometimes with further possibilities flagged in parentheses).

The great methodological insight which Stevenson's paper exploits is that we can gain clarity about the nature of moral thought, opinion, and judgment by better understanding the underlying character of moral disagreement. This strategy turns not on the frequency or pervasiveness of moral disagreement but on consideration of its very *nature*; it is a direct moral-psychological route to a moral psychology conclusion. If the argument succeeds, we can motivate non-cognitivism purely from a close examination of moral thought itself, without relying on (e.g.) a controversial metaphysical thesis about what is absent from the universe, a controversial semantic thesis about meaning or lack thereof, or an abductive argument proposing an explanation for certain empirical data. If Stevenson's general form of argument works we would have reason to incline toward a non-cognitivist construal of moral thought, opinion, and judgment even without any of these.

The argument that I see in Stevenson starts from the supposition that moral disagreement exists and invites us to ponder its character. The first hypothesis to consider is that moral disagreement fits a pattern already familiar to us from countless cases of disagreement lacking any moral (normative, evaluative) aspect. If Adam thinks the capital of Belarus is Minsk and Betty thinks the capital of Belarus is Pinsk, they disagree in the familiar way I have in mind. Stevenson calls this "disagreement in belief" and says it is characterized by "an opposition of beliefs, both of which cannot be true." This kind of disagreement, then, is constituted by the parties' having opposed *cognitive* states. Let us call it "type 1 disagreement."

Now suppose that (genuine) disagreement were a *unitary* phenomenon, in the sense that *all* genuine disagreement fit the above pattern.

The Narrow View: All genuine disagreement is type 1 disagreement.

The Narrow View would clearly not favor a non-cognitivist understanding of moral thought, opinion, or judgment. Suppose that Carol thinks it is morally wrong to eat meat and Dario thinks it's not morally wrong to eat meat. Suppose further that this should be counted as a genuine disagreement, not merely an apparent or verbal disagreement. Then by *The Narrow View*, the disagreement in question consists in an opposition of *beliefs*; but since belief is the archetype of a *cognitive* state, this would only tend to support the *cognitivist* idea that moral thought, opinion, and judgment are cognitive or doxastic phenomena. As we have seen, that is exactly the thesis that the non-cognitivist wishes to deny.

It will be important, then, for Stevenson to argue that disagreement is *not* a unitary phenomenon: that not all genuine disagreement takes the above form. In order to convince us of this, the non-cognitivist will need to demonstrate that there is some *other* species of genuine disagreement which does not conform to the profile of type 1 disagreement. She will need to convince us that the following is more convincing than *The Narrow Claim*:

[199] I make a simplifying assumption here and call "conative" any attitude that is not cognitive or doxastic.

The Existential Claim: There is at least one species or subcategory

of genuine disagreement distinct from type 1 disagreement.

The best way for her to demonstrate this would presumably be by *exhibiting* this putative distinct type of disagreement. Ideally it would be one familiar to us even outside the context of moral (normative, evaluative) thought, so that we can be sure it exists even if we set aside the category which is in dispute.

Suppose the non-cognitivist *does* convince us of the requisite existential claim: that there is a subcategory of genuine disagreement which does not fit the profile of what Stevenson called "disagreement in belief." Let us call the assumed witness for *The Existential Claim*—that is, the new, distinct species of disagreement—*type 2 disagreement*. The broad non-cognitivist gambit will then be to argue along the following lines:

[*Assumptions:* Carol and Dario genuinely disagree; and this is a moral disagreement.]

Classificatory Claim: Carol and Dario's disagreement is a type 2 disagreement.

Generalized Classificatory Claim: Moral disagreement falls under type 2 disagreement.

.

.

.

.

Conclusion: Moral opinions (&c.) are not cognitive, or doxastic, states or attitudes;

rather, they are *conative* states or attitudes.

Even without having yet filled in the ellipsis we can see that *Generalized Classificatory Claim* must somehow *support Conclusion* if the argument is to work. What must be true of type 2 disagreement in order for that to be the case? In order to play the necessary role in the argument, this assumed second type of disagreement must (it seems) involve *conative* attitudes. Without that, it would be a *non sequitur* to conclude from the supposition that moral disagreement instantiates type 2 disagreement that moral opinions (&c.) are conative attitudes. If, on the other hand, type 2 disagreement constitutively involves opposed *conative* attitudes, then *Generalized Classificatory Claim* really does provide support for viewing moral opinions (&c.) as built out of conative states.

Two paragraphs back, we introduced the term "type 2 disagreement" simply to indicate the (assumed) witness to *The Existential Claim*, without making any suppositions about its nature. The question, "Is there such a thing as type 2 disagreement?" is thus strictly equivalent to the question "Is *The Existential Claim* true?" But as we have just seen, the mere truth of *The Existential Claim* does not get the non-cognitivist's argument going unless we assume something further about type 2 disagreement, namely that it constitutively involves opposed conative attitudes. Moreover, it must be plausible to classify moral

disagreement under type 2 disagreement, otherwise we lose the *Classificatory Claim* and the *Generalized Classificatory Claim* on which the argument depends.

These last observations give us two further conditions on what the non-cognitivist would need in order to make her case. We have a kind of "black box" which needs to be dropped into the argument in order to complete it: let us call that black box "disagreement in attitude," and summarize the characteristics it would need to possess. In order to fill in the blanks in the potentially powerful argument that I have credited to Stevenson, disagreement in attitude would need to be

a) a species or subcategory of genuine disagreement

b) distinct from type 1 disagreement,

c) constituted by certain conative attitudes of the parties, and

d) under which it is plausible to classify moral (normative, evaluative) disagreement.

On my functional understanding of disagreement in attitude, we should answer the question in my title in the affirmative only if there exists a phenomenon which satisfies all of these conditions.

I want to underline that only something satisfying *all* of a)-d) can do the necessary work in the argument I outlined above. If, *contra* a), so-called type 2 disagreement were not really *genuine* disagreement, then classifying moral disagreement under it would serve only to take moral disagreement out of the category of genuine disagreement, contrary to our assumptions. If, *contra* b), so-called type 2 disagreement were simply a subclass of type 1 disagreement, then *The Existential Claim* would not after all be vindicated. If, *contra* c), type 2 disagreement did not essentially involve conative attitudes, then classifying moral disagreement under it would do nothing to support a non-cognitivist moral psychology. And if (*contra d*)) it were not plausible to see moral disagreements as type 2 disagreements, then the existence of type 2 disagreement would be irrelevant to the task at hand, which is to better understand the nature of moral thought, opinion, and judgment.

We have in effect written up a job description for disagreement in attitude. Does anything satisfy it? *That* is the question I am asking in my title, and that I now wish to explore.

2. What is Genuine Disagreement?

We will need at least a schematic understanding of (genuine) disagreement in order to proceed. For the first requisite on type 2 disagreement is that it disprove *The Narrow View* in favor of *The Existential Claim*. This means it must constitute a) a form of genuine disagreement b) that does not follow the type 1 template. Is there such a thing? [Spoiler alert:] I think Stevenson was right and there is such a thing. (If there isn't, our "powerful argument for non-cognitivism" goes cold right away.) Indeed, I think Stevenson gave us some examples of it. But I think he mischaracterized the type of disagreement which those examples instantiate. In order to substantiate these claims, I'll first need to offer up my own reflections on what is essential to genuine disagreement.[200]

Let's start with some uncontroversial observations. Disagreement is constituted by the holding of a certain polyadic relation, DISAGREE; let's call the relata of this relation the *parties* to the disagreement. (These will usually be persons.) Disagreement entails difference: if A and B disagree, then A and B differ

[200] I draw here from some paragraphs of my "Conceptual Disagreement," forthcoming (January 2019) in *American Philosophical Quarterly* (special issue on *The Nature and Implications of Disagreement*, eds. Palmira and Stroud).

qualitatively in some respect. But—importantly—the converse of the above does not hold: a mere (qualitative) *difference* between two people does not entail a *disagreement* between them. Suppose Elias and Fernanda differ in many respects: Elias is short, Fernanda tall; Elias has green eyes, Fernanda brown; etc. We could continue more or less indefinitely in this vein without yet having established any *disagreement* between Elias and Fernanda.[201] What, then, marks a qualitative difference between two people as a *disagreement*?

Two key points to note in this connection are i) if two parties disagree, there is something *at issue* between them; and ii) two parties cannot "just plain" disagree: if they disagree, they disagree *about something*. (Compare: a person cannot "just plain" *want*: if it is true that she wants, she wants *something*.) A and B disagree, in other words, only if there is something at issue between A and B; some *x* such that A and B disagree *about x*. Let us call the witness to this existential claim the *object* of their disagreement. Formally put, the claim is that the dyadic relation DISAGREE (A, B) can hold only if there is some *o* for which the *triadic* relation DISAGREE ABOUT (A, B, *o*) is satisfied:

Lemma 1. DISAGREE (A, B) $\supset \exists o$ (DISAGREE ABOUT (A, B, *o*))

Does this observation help to carve out genuine disagreements from mere differences? On a minimal interpretation of "about" as simply meaning *with respect to*, it might seem not to. Elias and Fernanda, for instance, differ (qualitatively) *with respect to* height, eye color, and mood. (Indeed, it sounds analytic to say that whenever two things differ (qualitatively), they differ *in some respect*, which one might think means: *with respect to something*.) But as we already noted, these facts about height, eye color, and mood do not constitute a *disagreement* between Elias and Fernanda. DISAGREE ABOUT (A, B, *o*) is not satisfied, then, merely because A and B differ (qualitatively) *in some respect o*, that is, *with respect to* some *o*.

Rather, as we said, when two parties *disagree about* something, there is something *at issue* between them. (There is nothing *at issue* between Elias and Fernanda in virtue of their different eye colors.) What *kind* of thing could be "at issue" between two parties? I submit that an ontological category well suited to fulfill this role is that of the *question*: as a general matter, when A and B disagree, a certain *question* is at issue between them. What then makes a particular question Q count as "at issue" between A and B? Q is certainly not "at issue" between A and B if A and B both answer Q in the same way, for instance. Indeed, the clearest cases of disagreement are the ones in which A and B take opposite positions on Q, i.e., they return opposite answers to Q. To sum up:

There is something *at issue* between A and B $\supset$

There is a question on which A and B *take opposed positions*

This formula would explain why there is nothing at issue between Elias and Fernanda in virtue of the brown eyes of the one and the green eyes of the other: to have brown eyes, or green, is not to *take a position* on any question. It would also explain why their differing *moods* do not constitute a disagreement (unless we decide that to be in a certain mood *is* to take a position on a question). In sum, the "aboutness" relevant to the triadic DISAGREE ABOUT relation is not merely the "with respect to" kind

[201] Note that this point holds even if we bring differences in *mental or psychological state*—not just in physical characteristics—into the picture. Suppose Elias is currently bored and Fernanda is currently excited. While they are currently in qualitatively different mental or psychological states, this does not by itself constitute a *disagreement* between them.

of aboutness; rather, it is the "taking a position on a question" kind of aboutness. This yields the following lemmas (where *q* ranges over questions):

Lemma 2. DISAGREE ABOUT (A, B, *q*) ⊃ A and B take opposed positions on *q*

Lemma 3. DISAGREE (A, B) ⊃ ∃*q* (A and B take opposed positions on *q*)

(*from Lemmas 1 and 2*)

We have arrived at the view that two parties genuinely disagree only if there is some *o*—the *object* of their disagreement—on which they take opposed positions, where *o* is a question. *That* is what makes a *difference* between A and B constitute a *disagreement*. As we continue to explore the prospects for "disagreement in attitude," this will be an important constraint on the scope of genuine disagreement. Now one might think we don't really understand this formula until we know what makes two positions count as *opposed*. I think this question is less important than the one I will especially probe, namely what kinds of questions are suited to be the *object* of a disagreement. It will be sufficient for present purposes to suppose that A and B's positions on some object *o* are *opposed* iff one of A and B *endorses o* and the other *rejects* or *repudiates o*.[202] So A and B will count as having opposed positions on a *question Q* when there is some possible answer to *Q* which one of them endorses and the other rejects. This will be the case, notably, when one of them returns an affirmative answer to *Q* and the other a negative answer to *Q*.

We need to make one more distinction before turning to the question of whether there are multiple distinct *kinds* of genuine disagreement, as *The Existential Claim* maintains. "Disagree" is ambiguous as between an activity and a state. In the *activity* sense, disagreeing is something you *do*, e.g. by objecting to or taking issue with something someone said. It is a type of speech act. In the *state* sense, however, you can disagree with someone without ever saying anything. You and she disagree if your *psychological states* stand in a certain relation, even if you scrupulously avoid ever speaking of the matter. (Indeed, that you both know you disagree about the matter might be the reason why you take care not to speak of it.) Now in the context of an argument meant to establish a non-cognitive moral psychology, it is clearly the *state* or *psychological* sense of disagreement that will be relevant. For the rest of the paper, it should be understood that we are always talking about disagreement in this psychological sense.

3. Varieties of Disagreement

As I mentioned earlier, I think Stevenson was right that there are at least two distinct varieties of (psychological) disagreement. I will exploit the fact that different kinds of things are suited to be *objects* of disagreement in order to establish this. My guiding methodological supposition will be that we can individuate types or kinds of disagreement by the types or kinds under which their respective *objects* fall.

Let us consider what we called "type 1 disagreement": the plain-vanilla kind with which we are all familiar. We supposed that Adam thinks the capital of Belarus is Minsk, Betty thinks the capital of Belarus is Pinsk, and moreover that this constitutes a genuine disagreement between them. It follows from this last supposition that there is something at issue between Adam and Betty, something about which

[202] As we shall see later, what precisely constitutes endorsement and rejection will vary with the *kind* of object of disagreement at issue.

they disagree. By our lemmas, there must be some object *o* on which they take opposed positions: there must be some question which one of them answers in the affirmative and the other in the negative—something which one of them endorses and the other rejects.

It is not hard to exhibit such an object or such a question. Adam and Betty disagree, for instance, about the following question:

Is Minsk the capital of Belarus?

or

whether Minsk is the capital of Belarus.

To each of these questions, one of them returns an affirmative answer, the other a negative answer. We could also say that Adam and Betty take opposite positions on the following proposition:

Minsk is the capital of Belarus.

One endorses it—which in this context means *thinks it's true*—while the other rejects it, i.e., thinks it's false.

What *kind* of thing is serving as the object of the disagreement in this case? –A proposition, or—in the interrogative forms—(what I'll call) a *propositional question*. I think we can generalize from this instance and say that when there is a proposition which one party thinks is true and the other thinks is false, we have a type 1 disagreement between the parties. Type 1 disagreement, in sum, is in essence *disagreement about whether a certain proposition is true*. Let's make that apparent by giving type 1 disagreement a new and more descriptive name: *propositional disagreement.* The following table summarizes what we have so far:

	what is at issue between the parties?	**to what type of question do the parties return opposite answers?**	**what is being endorsed or rejected?**	**what does the relevant endorsement or rejection consist in?**
propositional disagreement	whether *p* (whether *p* is true), for some proposition *p*	a propositional question: *p*? (is *p* true?)	the truth of a proposition	a cognitive or doxastic attitude

According to *The Narrow View*, we're done: the species of disagreement that we've just characterized constitutes the entire genus. *All* disagreement is disagreement about whether a certain proposition is true; the truth of some proposition or other is the only possible object of any disagreement deserving the name. At this juncture, however, it will be instructive to look at some of Stevenson's examples. He, and I, think they suggest we are *not* done: some genuine and perfectly familiar kinds of disagreements have a different character than the Minsk/Pinsk one.

We have, for instance:

> Two men are planning to have dinner together. One wants to eat at a restaurant that the other doesn't like. Temporarily, then, the men cannot "agree" on where to dine (Stevenson, p. 2).

We also have Mr. and Mrs. Smith, who "disagree … about whom to invite to their party" (p. 2). Mr. Smith proposes to invite his poker buddies, whereas Mrs. Smith wishes to have only members of the DAR darken her door. Finally, we have union and management (p. 4), who are at loggerheads concerning the upcoming collective agreement which they are now negotiating.[203] The union is militating for significant wage increases, but management is not willing to move wages above present levels.

All three of these dyads seem to be *in disagreement*, in a perfectly good, and indeed familiar, sense. (They are certainly not currently *in agreement*, at any rate.) But what *kind* of disagreement are they in? I propose that we pick out the *type* of dispute in which they are engaged by identifying the *object* of each of these disagreements. What are the parties disagreeing *about*? In the case of the two friends, it is

where to dine tonight;

for Mr. and Mrs. Smith, it is

whom to invite to their party;

and for union and management, it is

what terms to put in the next contract.

The above questions are not (or so I submit) primarily about whether a certain proposition is true. They are questions of a different kind: *practical* questions, of the sort that face deliberating agents. These are questions whose answers are not yet fixed, and which will remain open until the deliberating agent decides or acts. Indeed, we could say these questions will be settled *by* the agent's decision or action. It is up to the agent, or agent*s*, to *determine* the answer to a practical question facing them; by which I mean that it is up to them to *fix* the answer, not merely to determine (in the sense of *coming to a view about*) whether a certain proposition is true. Rather, the agent can *make it the case*, through her action or decision, that a certain answer to the question facing her is correct.

We noted earlier that in a type 1 disagreement what is at issue can be expressed as a *wh*-question, namely *whether* p *is true* (for some proposition *p*). In the cases just above, too, we naturally expressed what is at issue with a *wh*-question, but this time with an *infinitival* complement rather than a propositional one. This infinitival construction is typical of practical questions, as befits the fact that the agent trying to resolve a practical question is choosing among potential *actions*. If the schematic form of a propositional question is

whether *p* (or: whether *p* is true),

for some proposition *p*, the schematic form of a *practical* question is

whether to φ,

[203] Stevenson does not explicitly say that they are in the midst of negotiating the next contract, but I don't think he would disagree with that reading of the case, and it is important to my own analysis that this is the context.

for some action φ.

Moreover, we can express a practical question in an interrogative sentence, just as we can a propositional question like "is Minsk the capital of Belarus?" English has a handy way of flagging distinctively practical questions, namely the auxiliary verb *shall*. Speculating about a couple on the other side of the ballroom, you and I might ask ourselves,

Will they dance?

But the parties concerned have a different, non-predictive question on the table, namely,

Shall we dance?

This, for them, is a *practical* question, to be resolved by their deciding to dance, or not. We deploy *theoretical* reasoning to answer propositional questions like "Will they dance?", but *practical* reasoning to answer questions like "Shall we dance?"

A practical question, then, is a question that we engage in practical deliberation in order to settle. We are most familiar with such questions from the first-person-singular case, the "I" case. Whenever we deliberate about what to do, we are entertaining and seeking to resolve a distinctively practical question which faces us. ("Shall I take the job?") Except in the very rare case in which a third party is in a position to determine what *I* shall do, a third party cannot even ask the question which occupies me as an agent and which is mine to resolve: a question can only be a genuinely practical one for *S* if *S* is in a position to fix the answer to it through her action or decision. So practical questions are by their very nature first-personal. But this restriction does not exclude the case which will particularly interest us, namely practical questions in the first person *plural*. Only I have dominion over what I shall do. But *two* parties can have jurisdiction over what *we* shall do. Shall we dance? (It takes two to tango, after all.) Shall we buy this house (a married couple might wonder)? Shall we write a paper together (two colleagues might ask)? Shall we put that poster here (ask the college roommates)? Such questions arise when *we* need to decide what *we* shall do.

The dyads in Stevenson's examples are grappling with questions of this kind. Shall we go to the Chinese restaurant tonight? Shall we invite the poker buddies to the party? Shall we provide for increased wages in the next contract? Stevenson's protagonists need to decide the answers to those questions. So far, though, no decisions are forthcoming, as the parties have not yet reached agreement. The eventual collective agreement will have to be signed by both sides; but right now neither side is willing to sign the text the other is proposing. In a perfectly good sense, union and management disagree about what terms the next contract shall contain.

I think, with Stevenson, that *de nobis* practical questions of the kind that these dyads are facing make a second type of disagreement possible. I propose to call it *practical disagreement*. Practical disagreement, I maintain, instances *The Existential Claim*: it is a) a type of genuine disagreement, but b) a different variety than type 1 disagreement. Let me take the above points in turn. On a): we said earlier (pp. 10-11) that genuine disagreement is marked by the parties' taking opposed positions on a common object of disagreement *o*. We characterized "opposed positions on *o*" as one party's endorsing, and the other rejecting or repudiating, *o*, and we noted that in a genuine disagreement there is some question which one party answers in the affirmative and the other party answers in the negative. All of these conditions are present in the cases we have just considered. There is indeed a question to which the parties in the union negotiation (for instance) return opposite answers, namely the practical question

Shall we raise wages in the next contract?

One of the parties endorses, and the other repudiates, a possible answer to that question:

We shall not raise wages in the next contract.

I will henceforth call a potential answer to a practical question—linguistically flagged, as above, with the auxiliary verb "shall"—a *plan*. Our disputants take opposed positions on the above plan: one is for it, the other against it. We thus have our object of disagreement, *o*.

These cases, then, constitute genuine disagreements: they fit the schematic profile of disagreement which I proposed earlier. However, it is important to note that the *interpretation* of that schema's provisions varies with the category to which *o* belongs, yielding different species of the genus. In type 1 disagreements, *o* is a proposition or propositional question, and the parties' opposed positions on *o* consist in opposite *doxastic stances* on the truth of that proposition. But that is not the right description of the present cases. In a *practical* disagreement, while it is correct to say that A dissents from B's position, A's objection is not (as in propositional disagreement) that B thinks something which is false. It is rather that A does not *support*, in a more conative sense, B's position. So while we have a genuine disagreement here, it is not a disagreement of the same kind. If this is right, disagreement comes in at least two varieties; *The Existential Claim* is vindicated.

Let us sum up what we have learned. When the question at issue between two parties is a *practical* one, the object which one party endorses and the other rejects is a *de nobis* plan, not an ordinary proposition. And that shift in object yields a change in the kinds of attitudes which properly constitute endorsement or rejection of an object of that type. When oriented toward *plans* rather than regular propositions, endorsement and rejection need to be understood as conative, not doxastic, attitudes. Our taxonomic table, then, has now grown:

	what is at issue between the parties?	**to what type of question do the parties return opposite answers?**	**what is being endorsed or rejected?**	**what does the relevant endorsement or rejection consist in?**
propositional disagreement	whether *p* (whether *p* is true), for some proposition *p*	a propositional question: *p*? (is *p* true?)	the truth of a proposition	a cognitive or doxastic attitude
practical disagreement	whether to φ, for some *de nobis* action φ	a *de nobis* practical question: shall we φ?	a *de nobis* plan	a conative attitude

4. Whither Disagreement in Attitude?

I have expressed the opinion that Stevenson was right to think there is a second distinct type of disagreement which does not follow the familiar propositional model and which instead constitutively involves conative attitudes. Things seem to be looking good, then, for the existence of disagreement in attitude: I seem to have conceded that there is a phenomenon satisfying a), b), and c) of the job description which we wrote up for that concept (see p. 7). But despite my appreciation of Stevenson's insight concerning these examples, I don't think they secure disagreement in attitude as he conceived of it. I want now to defend the following theses:

1) Stevenson overgeneralized from these examples, supposing they established more than they do.

2) There is no disagreement in attitude beyond practical disagreement. Any supposed disagreement in attitude that is outside the boundaries of the latter is merely what might be called *disattunement*, not true disagreement.

3) If 2) is true, the *Generalized Classificatory Claim* is implausible, and our best candidate for satisfying a) through c) of the job description for disagreement in attitude fails to satisfy d).

What did Stevenson take himself to have demonstrated with the examples we have discussed? He thought they showed the existence of a type of disagreement which

> occurs when Mr. A has a favorable attitude to something, when Mr. B has an unfavorable or less favorable attitude to it, and when neither is content to let the other's attitude remain unchanged. The term "attitude" … designates any psychological disposition of being *for* or *against* something (pp. 1-2).

Now Stevenson, of course, used the term "disagreement in attitude" to refer to the phenomenon picked out by the passage just quoted, and whose existence he took his examples to demonstrate. But to avoid confusion—since we are operating with a *functional* understanding of disagreement in attitude—we should give his phenomenon a different name. (It would beg the question to suppose that the phenomenon he defines satisfies all the requisites to be disagreement in attitude in our functional sense.) Let us use the label "attitudinal opposition" to denote the phenomenon he identifies in the passage above.

Stevenson's use of the term "when" in this passage suggests a *sufficient* condition for attitudinal opposition. Setting aside the last clause of the account, which is typically viewed as a puzzling addition,[204] the claim we will extract from the passage is that

A and B instantiate *attitudinal opposition*

if there is some *x* such that A is for *x* and B is against *x*.[205]

[204] It has always been viewed as curious that Stevenson added that condition about not being content to let the other person's position remain unchanged to his specifications of both "disagreement in attitude" and, especially, "disagreement in belief," where it appears particularly inapposite. The subsequent literature has not followed him in thinking some such condition requisite for genuine disagreement, at least not in the psychological or "state" sense (see p. 12). For this reason I will engage with the simplified version of his view that has no such further condition. I believe the points I will make against that view would hold even if we added this further condition, however. See note 9.

[205] There are alternative ways we might phrase the "such that" clause: for instance, drawing on the passage we quoted earlier, we might say "…such that A has a favorable attitude and B an unfavorable attitude toward *x*," or (using more modern terminology) "…such that A has a pro-attitude, and B a con-attitude, toward *x*." If these differ materially from the construction in the main text, the differences will not matter to our treatment.

Attitudinal opposition, so defined, is crucially broader than practical disagreement as we previously defined it, for which the corresponding claim would be

A and B instantiate *practical disagreement*

iff there is some ***de nobis* plan *P*** such that

A is for *P* and B is against *P*.

The Stevensonian formula for attitudinal opposition does not insist that the parties' opposed attitudes be directed specifically at a *de nobis* plan or practical question in order to constitute a disagreement between them. Instead, it suggests no restrictions on the scope of *x* whatsoever. I think this is a significant—and mistaken—omission which makes most attitudinal oppositiobn not disagreement at all. I now pass to my claim 2): the simple fact that one person is *for* (or has a pro-attitude toward) something which another person is *against* (or has a con-attitude toward) does not by itself constitute a *disagreement* between them.

Suppose we let the universe of discourse for potential objects of disagreement range unrestrictedly over anything toward which a pro- or con-attitude can be directed. As I have emphasized, this will be a significantly broader universe of discourse than that for potential objects of *practical* disagreement, which are limited to *de nobis* practical questions and their potential answers, namely *de nobis* plans. The larger universe of discourse will contain, in particular, items that fall under two categories absent from the latter, smaller set: i) things that are not practical questions or plans at all, and ii) practical questions and plans which are first-person-singular (*de se*) rather than *de nobis*. I believe neither of these classes of things can support genuine disagreement as opposed to mere difference or disattunement.

My case will be merely anecdotal and suggestive, but consider some examples. My husband, for instance, dislikes Japanese maple trees; I rather like them. We differ qualitatively in this respect; we are "not on the same page" when it comes to Japanese maple trees. But do we *disagree*?[206] We have opposite attitudes toward one and the same thing, certainly. But can a *tree*—as opposed to a proposition about a tree, or a plan involving a tree—be a proper object of disagreement? Now one might object that this example is unfair, because by "attitude" we should mean *propositional attitude*, i.e., an attitude directed toward a proposition, and a tree is not a proposition. Point taken: let us consider then *past events*, toward which we can certainly have pro- and con-attitudes which are naturally expressed using "that"-clauses.

Suppose I am glad that Caesar crossed the Rubicon, or that the Red Sox won the World Series this year. You, on the other hand, regret both of these bitterly. We certainly have *contrasting attitudes* toward one and the same event, or state of affairs: our respective attitudes are without a doubt *opposed* to each other in that sense. But, once again, does this mean we *disagree*? By our *Lemmas*, we disagree only if there is something that we disagree *about*. What exactly could that be? What, if anything, is *at issue* between us, simply in virtue of these contrasting attitudes? We react differently to these events, yes. But what is the question, if any, on which we come down on opposite sides? I do not think one can be identified. By our *Lemmas*, there needs to be something that each of us is *taking a position* on; but it is

[206] Do we disagree even if his dislike of Japanese maples bugs me for some unaccountable reason, and I'm always trying to get him to change how he feels? I *don't* have in mind the case in which something hangs on it, such as whether to cut down the Japanese maple in the backyard.

not clear that the mere fact of reacting positively or negatively to a past event rises to the level of taking a position on anything.

Notice that it would be dialectically improper to propose, as the object of our putative disagreement, an issue like this:

whether it is a good thing that the Red Sox won the World Series.

This would be unfair because our *aim*, precisely, is to understand better what it is to think such a thought as "it is a good thing that the Red Sox won the World Series": in particular, whether this consists in a doxastic or a conative state. Such thoughts are, in the present dialectical situation, *disputed territory*: we are trying to decide whether to file them with the (undisputed) *beliefs*, or the (undisputed) *conative states*. If we are to classify our contrasting attitudes toward the Red Sox' victory as a disagreement, we need to identify something that is *unquestionably* eligible to be an object of disagreement—a proposition, or a plan—as being what is at issue between us.

I have suggested that there are various *categories of thing* toward which it is possible for us to have contrasting (conative) attitudes without thereby being in disagreement. In particular, I have suggested that things other than practical questions and the potential answers to them (plans) do not support genuine disagreement of this conative kind. The same holds, I now argue, for first-person-*singular* practical questions and plans. Suppose you and I are two strangers on a plane. We each have a plan for when we disembark: I plan to take the subway into the city, whereas you plan to take a taxi. We have contrasting plans: the contents of our respective plans differ and indeed are incompatible (in at least the sense that it would be incoherent for a single person to have *both* plans). Do we disagree? No, we just have different plans. There is nothing at issue between us: there is no common practical question on which we are butting heads. It is true that we return opposite answers to the practical question

Shall I take a taxi into the city upon landing?,

but we no more disagree in virtue of this than we do when I endorse and you reject the statement

I am a Canadian citizen.

The situation changes if the two of us are a married couple and see it instead as *our* decision how *we* will get downtown after the plane lands. *Now* there is a common practical question on the table, and on which we might take opposed positions:

Shall we take a taxi into the city upon landing?

This scenario supports the possibility of genuine (practical) disagreement in a way that contrasting *de se* plans did not. This point can be obscured by one way of expressing what is at issue between the spouses, namely

whether to take a taxi into the city upon landing.

When we put the object of our disagreement *this* way, it might seem equally available to be an object of disagreement in the strangers on a plane case. After all (you might reason), the first traveler answers that question in the negative, whereas the second answers it in the affirmative. But the appearance of univocity here is merely an artifact of the unclarity induced by the infinitival phrase "whether to φ." The infinitival complement leaves the imagined agent of the action unstated; for maximum clarity we ought to

disambiguate "whether to $\varphi_{de\ se}$" from "whether to $\varphi_{de\ nobis}$." But the construction using "shall" (see above) does this more neatly.

I have been arguing that what we have called "attitudinal opposition" (as defined on p. 23) is not really disagreement, except in the special case in which it is practical disagreement. That is, I have been urging that only *de nobis* practical questions and plans are appropriate objects of disagreement of this conative kind. To vindicate a claim that two parties genuinely disagree, then, we must be able to identify either a proposition (if type 1) or a *de nobis* practical question or plan (if type 2) as the thing about which they disagree. This point has, I believe, serious implications for the non-cognitivist argument with which we began. In particular, it makes it much more difficult for the non-cognitivist to motivate the idea that moral (or ethical, or normative) disagreement is type 2 disagreement. For corresponding to every genuine moral disagreement between two parties, there would have to be a *practical* question facing the parties *as a "we"* to which they return opposed answers. Moral thought would be simply a subset of *de nobis* practical thought, with a scope no broader than the scope of the latter.

Can we identify such a question for every, or indeed any, moral disagreement? Recall Carol, who thinks it's wrong to eat meat, and Dario, who thinks it isn't. A *practical* question, recall, is distinguished by the following unusual feature: it is within the agent's power to *make* one answer to it the correct one. For the agent's decision or action will *fix* the answer to the question. When *we* face such a question, *our* decision or action will be what settles it, as we saw with Stevenson's examples. Yet it is not at all obvious what is within Carol and Dario's power in this way when it comes to eating meat. Why are Carol and Dario in a position to settle or fix anything in this regard?[207] They cannot make it the case that it *is* wrong to eat meat, for instance, even if that were a kosher state of affairs (which the non-cognitivist will of course resist). Their situation is not analogous to that of the Members of Parliament, for example, who *are* in a position to make it the case, through their decision, that it is illegal to eat meat. Carol and Dario have no legislative power over moral wrongness, were there such a thing.

They do, as individuals, have legislative power over *some* things. Carol can decide, for instance, whether to eat meat herself; and Dario can do the same with respect to himself. But these are *de se* practical questions facing each of them separately, just as in our case of the strangers on a plane. I would say the same for questions like these:

whether to blame people who eat meat

whether to discourage people from eating meat

whether to internalize norms that forbid eating meat

whether to try to stop wanting to eat meat

and so on. These are at best practical questions facing each of Carol and Dario individually; and we argued above that when two parties return opposite answers to a *de se* practical question, that is merely a difference between them, not a disagreement. It seems we have failed to locate a type 2 disagreement with which to identify the moral disagreement between Carol and Dario.

[207] If they were a married couple, they could together settle *some* questions about eating meat, such as: Shall we serve meat in our house? But surely moral disagreements are not limited to spouses.

Conclusion

Disagreement in attitude needed to be many things in order to do the job that Stevenson—and, I believe, those in the Stevensonian tradition—wanted it to do. It had to be, first of all, a genuine species of disagreement, distinct from disagreement about whether a certain proposition is true. It had to be constituted by conative, not doxastic, attitudes. And lastly, moral (or normative, or evaluative) disagreement had to exemplify it. I have argued that we have not found a solution for this equation. Practical disagreement, I argued, is a genuine form of disagreement; and it is indeed constituted by conative attitudes. But it is *not* plausible to see a practical disagreement in every moral disagreement. Stevenson's favored candidate for disagreement in attitude was apparently "attitudinal opposition," as I termed it, which consists in one party's being *for* something that the other party is *against*. While it might be plausible to say that this is present whenever two people morally disagree, it is not, alas, a species of disagreement at all. Our equation remains, so far, unsolved.

It might be said that things have progressed a great deal since Stevenson's day. Perhaps subsequent developments in non-cognitivism have pointed toward a more satisfactory candidate for disagreement in attitude than those I have discussed. It might be hoped, for instance, that we could produce a better candidate by limiting Stevenson's coarse-grained conception of conative disagreement to a more specific *kind* of conative attitude. Perhaps being (respectively) generically *for* and generically *against* something is not enough to constitute a disagreement, but (say) *approving* of something vs. *disapproving* of it does succeed in taking us out of the realm of mere qualitative difference. This strategy, however, does not engage with a point I have tried to stress, namely the key role of the *object* of a disagreement in elevating a mere contrast of attitudes into something's actually *being at issue* between the parties. Without an appropriate object, we lack genuine disagreement; fiddling with the particular attitudes involved doesn't seem liable to change this.

A second line of attack might be to revisit the question of whether we can see a form of practical disagreement in all (genuine) moral disagreements. Perhaps, for instance, modelling moral (normative, evaluative) thought as commitment to *hyperplans* would allow us to place the former under the umbrella of practical disagreement after all. Contingency planning, it might be said, is a recognized part of practical thought; so if an "'ought' thought" in fact consists in an elaborate contingency plan, we are right to classify divergent normative views under type 2 disagreement. As against this hope I would want to underline once again the centrality of *de nobis* practical thought for genuine disagreement. However complicated it may be, a *de se* plan intended to guide *my* action affords me only a way of contrasting with—not disagreeing with—my neighbor.

Speaker Bios

Gregory Antill's research lies at the intersection of epistemology, ethics, and the philosophy of mind. Much of his current research focuses on the ways in which a subject's agency as both a believer and a moral actor can be effected by how the subject is situated in their environment. Gregory completed his Ph.D. at UCLA, and is currently a Visiting Assistant Professor at Claremont McKenna College.

Saba Bazargan-Forward is Associate Professor of Philosophy at the University of California, San Diego. His work primarily in normative ethics, with a focus on the morality of defensive violence, the morality of war, and complicity.

Andrew Flynn is a doctoral candidate at the University of California - Los Angeles. He is interested primarily in moral philosophy and its history, as well as philosophy of action, but he also has interests in related issues in philosophy of mind and epistemology. He is currently working on a dissertation on irony in moral philosophy.

Cami Koepke is a graduate student at UC San Diego working under the direction of Dana Nelkin, David Brink, and Manuel Vargas. He's especially interested in understanding how conditions like addiction might mitigate moral responsibility. Hes explore this issue in detail in his dissertation titled "Responsibility and Addiction: A Defense of a Control-Based Theory of Moral Responsibility." More broadly, he's interested in bioethics, philosophy of psychiatry, and philosophy of religion.

Max Kramer is a Ph. D. student in philosophy and cognitive science at the University of Arizona. He works at the intersection of philosophy of mind and ethics, with a particular emphasis on emotions, fellow-feeling, and social cognition. He also has interests in experimental philosophy and social ontology.

Lilian O'Brien received her PhD from Brown University and has taught at Vassar College, The College of William and Mary, and University College Cork, Ireland. She is University Researcher at the University of Helsinki.

Peter Railton is Gregory S. Kavka Distinguished University Professor at the University of Mchcigan. His main research has been in ethics and the philosophy of science, focusing especially on questions about the nature of objectivity, value, norms, and explanation. Recently, he has also begun working in aesthetics, moral psychology, and the theory of action. He has a special interest in the bearing of empirical research in psychology and evolutionary theory on these questions. A collection of some of his papers in ethics and meta-ethics, *Facts, Values, and Norms***,** appeared with Cambridge University Press in 2003.

Brian Reese is a doctoral candidate in the philosophy department at the University of Pennsylvania. He completed his masters work at the University of Oxford and his undergraduate work at the University of Michigan. Before undertaking graduate study, he worked for three years as an assistant editor at the Whitney Museum of American Art. Some of his editorial training is now being put to use at the *Archiv für Geschichte der Philosophie*, where he serves as an assistant managing editor.

Grant J. Rozeboom is an Assistant Professor of Business Administration-Ethics and Philosophy at St. Norbert College. He received his PhD from Stanford University in 2015. His research focuses on questions in moral philosophy, the philosophy of action, and applied ethics, and includes published work on the basis of equality, moral worth, and practical deliberation.

Sarah Stroud is Professor of Philosophy and Director of the Parr Center for Ethics at the University of North Carolina, Chapel Hill. She was previously Professor of Philosophy at McGill University, where she taught from 1993 to 2018. She holds degrees from Harvard (A.B.) and Princeton (Ph.D.). She works across central areas of moral philosophy, with a particular focus on foundational issues in moral psychology and moral theory and on the intersection of such issues with metaethics and the philosophy of action. She has published papers on such topics as partiality, moral demandingness and overridingness, lying and testimony, practical irrationality, and the moral implications and significance of personal relationships. She co-edited *Weakness of Will and Practical Irrationality* (OUP, 2003) and the *International Encyclopedia of Ethics* (Wiley-Blackwell, 2013).

Katherine Sweet is a doctoral student in the department of philosophy at Saint Louis University. Her interests are in Ethics, Epistemology, Philosophy of Religion, Political Philosophy

Nandi Theunissen is an associate professor at the University of Pittsburgh. She is interested in the nature of value, and her particular focus is the value of humanity. This is the subject of her book project, forthcoming with OUP later this year.

CHICAGO ATTRACTIONS

John Hancock Tower:
The best-kept secret in Chicago tourism is the Signature Lounge, located on the 96th floor of the Hancock Tower, 875 N. Michigan Ave. This bar/restaurant provides guests with a 360 degree view of Chicago and Lake Michigan for the price of a drink --there is no admission fee.

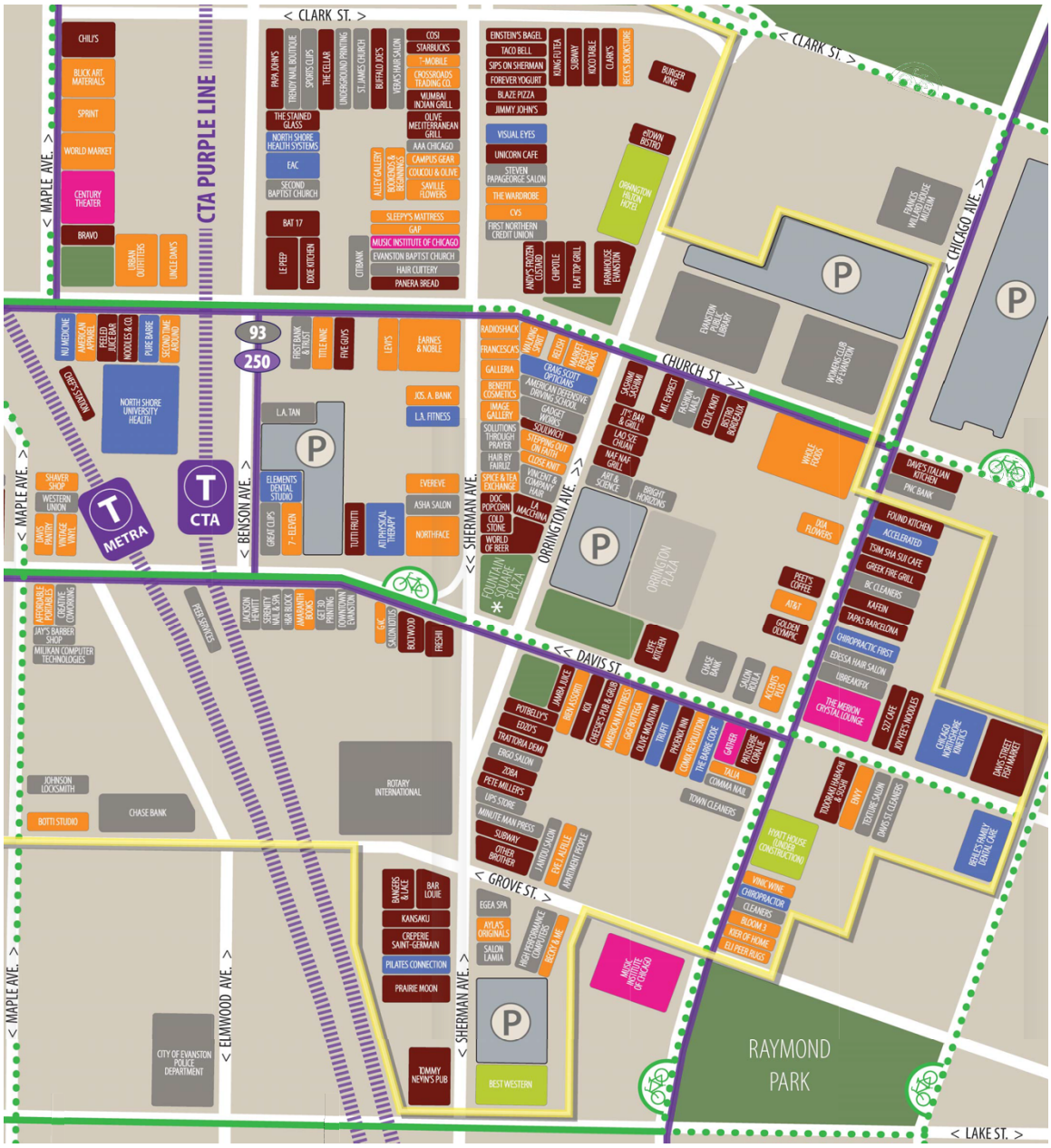
The Magnificent Mile:
Chosen as one of the ten great avenues of the world, the Mag Mile is located just north of the loop and is Chicago's most prestigious shopping district. Water Tower Place, a very large mall, is located at 835 N. Michigan Avenue. Walking south on Michigan Ave (or taking any of the many buses) you will end at the Wrigley Building down on the river (which you can follow into the loop and to Millenium Park and the Art Institute).

Chicago Architecture Foundation Boat Tour:
$44 for daytime crusies and $46 for nightime cruises, 90 minutes long. Dock location is southeast corner of the Michigan Avenue Bridge and Wacker Drive. Look for the blue awning marking the stairway entrance. You can buy tickets online.

Millennium Park:
Millennium Park is located in the heart of downtown Chicago. It is bordered by Michigan Avenue to the west, Columbus Drive to the east, Randolph Street to the north and Monroe Street to the south. This park is open daily from Sam to 11pm. Admission is free. Attractions include the enormous mirror-surfaced bean sculpture, the Cloud Gate bridge, the Crown Fountains, the outdoor ampitheater, and the Lurie Garden.

Shedd Aquarium:
Museum Hours: Weekdays: 9am-5pm & Weekends: 9am-6pm. Admission: $8 adults for aquarium only, $31 for all-access pass that includes Oceanarium, Wild Reef, Amazon Rising, the Caribbean Reef, Waters of the World, and others. To get to the museum, take the red line L to the Roosevelt stop and board a museum trolley or take the #12 bus.

The Field Museum:
Museum Hours: 9am-5pm. $38 for an all-access pass. Take the red line L to the Roosevelt stop and board a museum trolley or take the #12 bus.



EAT/DRINK

SHOP

BE ENTERTAINED

STAY