# Northwestern University

## Society for the Theory of Ethics and Politics

10th Annual Conference May 19-21, 2016

John Evans Alumni Center- 1800 Sheridan Rd., Evanston


Keynote Addresses:

"In Defense of Benevolence"

Nomy Arpaly, Brown

Friday, May 20, 4-6pm

"Standpoints and Freedom"

Pamela Hieronymi, UCLA

Saturday, May 21, 4-6pm

A SPECIAL THANK YOU TO THE GRADUATE SCHOOL AND THE WEINBERG COLEGE OF ARTS AND SCIENCES FOR THEIR GENEROSITY

**Table of Contents**

**NORTHWESTERN UNIVERSITY**
**SOCIETY FOR THE THEORY OF ETHICS AND POLITICS**

# A SPECIAL THANKS...

**CONTENT PROVIDERS:**
**OUR CONFERENCE SPEAKERS AND COMMENTATORS.**

**CONFERENCE ORGANIZERS:**
**KYLA EBELS-DUGGAN, RICHARD KRAUT, STEPHEN WHITE, ABIGAIL BRUXVOORT**

**FACULTY PAPER SELECTION:**
**KYLA EBELS-DUGGAN, RICHARD KRAUT, STEPHEN WHITE**

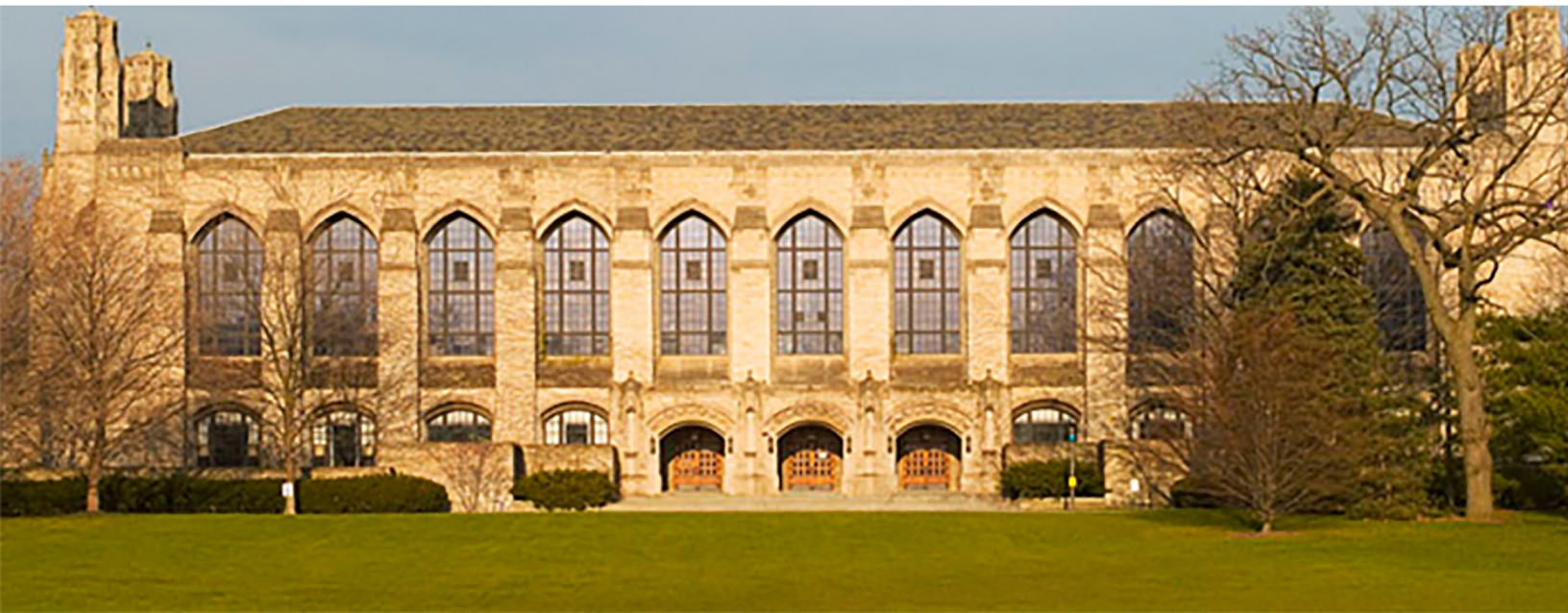
**ADMINISTRATIVE SUPPORT:**
**CRYSTAL FOSTER, JASMINE HATTEN, TRICIA LIU, NATHANIEL POLAND**

**SUBMISSIONS REVIEW COORDINATOR:**
**CHELSEA EGBERT**

**REVIEW SESSION ORGANIZER:**
**CARLOS PEREIRA DI SALVO**

**PROMOTION:**
**BLAZE MARPET, ANDY HULL**

**BUDGET/GRANT APPLICATIONS:**
**HAO LIANG, TAYLOR ROGERS**

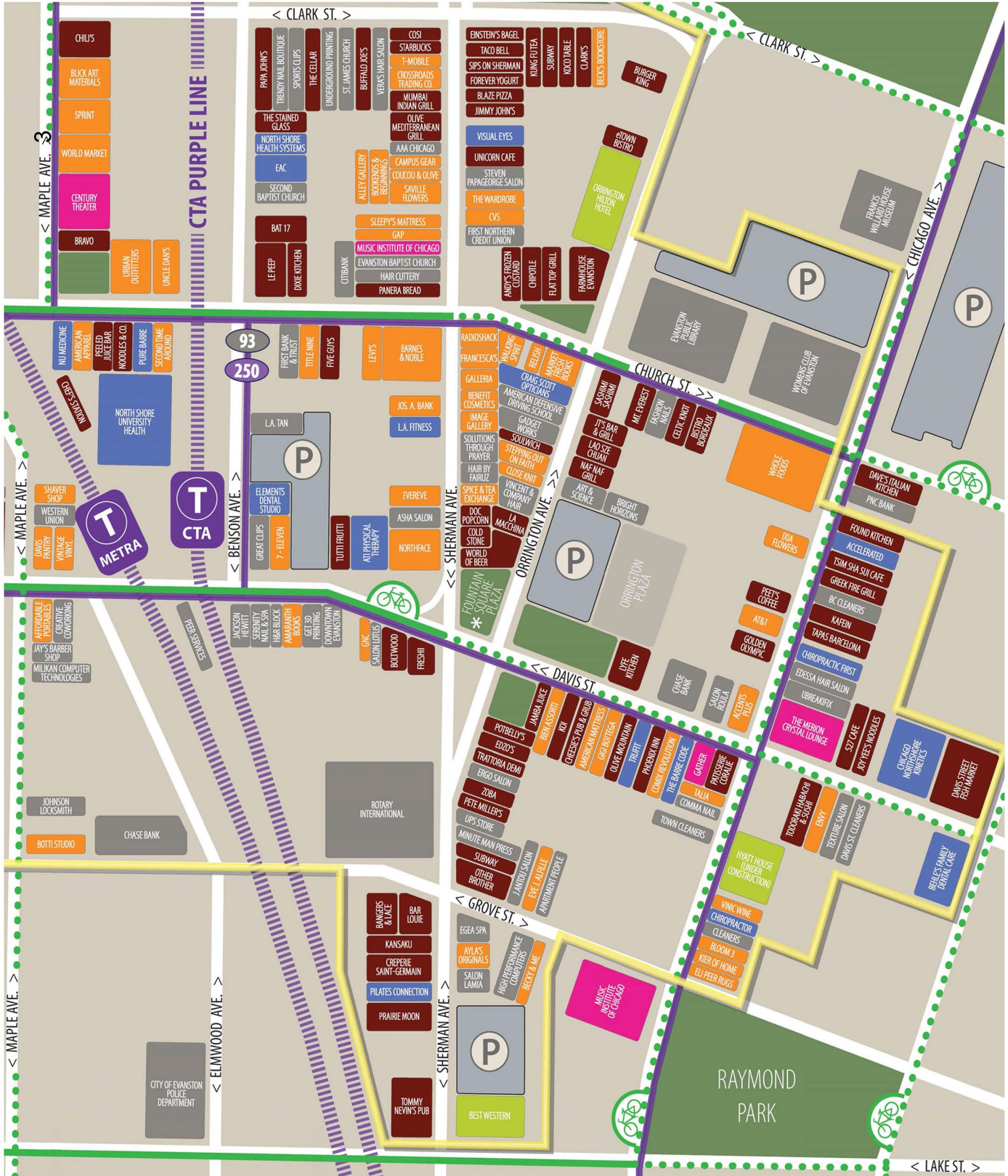


# RECEPTION INFORMATION

**All speakers and commentators are welcome to attend a buffet dinner on Friday evening following the afternoon session.**

**Location: John Evans Alumni Center**
**1800 Sheridan Road**
**Evanston, IL 60208**

**The reception will begin at approximately 6:00.**




## KEY

- PUBLIC PARKING
- TRAIN
- PACE BUS ROUTES (213)
- CTA BUS ROUTES (205)
- BIKE LANES
- BIKE ROUTES

EAT/DRINK

SHOP

BE ENTERTAINED

STAY

GET HEALTHY

TRAVEL

SERVICE

DOWNTOWN℠ EVANSTON

*Where Chicago and the North Shore Meet!*

Scan to discover !

# 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 8am 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.

# Northwestern University Society for the Theory of Ethics and Politics

10th Annual Conference
May 19-21, 2016
John Evans Alumni Center

## Thursday, May 19th 2016

Morning Session

9:00-10:25
***Neo-Republicanism Needs a Criterion of Reasonableness***
Kirun Sankaran (Brown)
Comments: Joshua Kissel (Northwestern)

10:35-12:00
***Hypocrisy and the Standing to Blame***
Coleen Macnamara (UC Riverside)
Comments: Rachel Fredericks (Ball State)

*Lunch*

Afternoon Session

2:15-3:40.
***The Nature of Blame and Our Reasons for Forgiveness***
David Beglin (UC Riverside)
Comments: Hao Liang (Northwestern)

3:50-5:15
***The Forgiven***
David Shoemaker (Tulane)
Comments: Heidi Giannini (Hope College)

*Dinner*

## Friday, May 20th 2016

Morning Session

9:00-10:25
***Post Hoc Ergo Propter Hoc: Some Benefits of Rationalization***
Jesse Summers (Duke)
Comments: Nir Ben Moshe (University of Illinois Urbana-Champagne)

10:35-12:00
***Trying is Good***
Zoë Johnson-King (University of Michigan)
Comments: Gretchen Ellefson (Northwestern)

*Lunch*

Afternoon Session

2:15-3:40
***The Value of Attachment***
Monique Wonderly (Princeton)
Comments: Yujia Song (Purdue)

3:50-5:45.
**Keynote Address:**
***In Praise of Benevolence***
Nomy Arpaly (Brown)
Comments: Kyla Ebels-Duggan (Northwestern)

*Reception – Everyone is invited*

**Saturday, May 21st 2016**

Morning Session

10:35-12:00
***The Kantian Conception of Obligation and the Directedness Constraint***
Aleksy Tarasenko-Struc (Harvard)
Comments: Abigail Bruxvoort (Northwestern)

*Lunch*

Afternoon Session

2:15-3:40.
***Do Reasons Expire?***
Berislav Marušić (Brandeis)
Comments: Benjamin Yelle (Mount Holyoke)

3:50-5:45
**Keynote Address:**
***Standpoints and Freedom***
Pamela Hieronymi (UCLA)
Comments: Maura Tumulty (Colgate)

*Dinner*

Kirun Sankaran
Brown University

**Bio**: Kirun Sankaran is a second year graduate student in philosophy at Brown who works primarily in political philosophy and its history. Before Brown, He finished his MA in 2014 at the University of Wisconsin-Milwaukee and his BA in 2012 at the Ohio State University. Outside of work, you can find him watching sports.

# Neo-Republicanism Needs A Criterion of Reasonableness[1]

Perhaps *the* fundamental commitment of liberal political philosophy is that the principles that govern the political life of a society must, because of their *coercive* character, be justifiable to citizens in some way. One of the lessons of Rawls' work after the "political turn" is that the fact of reasonable disagreement about the nature of justice means that acceptability to every citizen's point of view is neither possible nor morally advantageous. Thus, the set of citizens to whom justification is owed, and the reasons that can be brought into play to do the justificatory work must somehow be restricted. Liberals have recognized this for the past quarter-century, which has been spent in an extended argument about how best to understand that criterion.

Philip Pettit argues that his neo-Republican view[2] is a superior alternative to liberalism because its single political value of freedom, understood as non-domination, can justify the coercive state to the various reaches of a society characterized by widespread disagreement about justice. On his view, freedom as non-domination is analyzed in terms of interest-tracking. Because of this, I will argue, Pettit must restrict the set of interests to be tracked according to a substantive, normatively-weighty criterion. He must give an account of something analogous to the liberal notion of reasonableness. In order to illustrate this, I will rely on a set of cases I'll call "domination tradeoffs", which show that the value of non-domination cannot itself pick between possible ways of arranging the basic structure of society, and that a criterion of reasonableness is

[1] Thanks to David Estlund, Thomas Fisher, Nicholas Geiser, Robert Joynt, Charles Larmore, Daniel Layman, Ferris Lupino, Rebecca Millsop, Ryan Muldoon, Thomas Mulligan, Daniel Muñoz, and Michal ben Noah for helpful comments on this and previous drafts.

[2] Cf. Pettit (1997), (2012); For historical antecedents and a slightly different formulation see Skinner (1998).

needed to resolve this indeterminacy. This is a reason to believe that the neo-Republican view is at a disadvantage to the liberal one along this particular axis.[3]

## I. The Role of Reasonableness in the Rawlsian Liberal Tradition[4]

Starting in about 1980 with the publication of "Kantian Constructivism in Moral Theory," John Rawls' "political turn" spawned an entirely new literature in political philosophy. One central feature of this new "political liberal" project is its embrace of the Liberal Principle of Legitimacy:

> Our exercise of political power is fully proper only when it is exercised in accordance with a constitution the essentials of which all citizens as free and equal may reasonably be expected to endorse in the light of principles and ideals acceptable to their common human reason.[5]

On the political liberal picture, exercises of coercive authority are legitimate only when they can be justified to the coerced parties according to the principles and ideals "acceptable to their common human reason." Charles Larmore notes that, for Rawls, it follows from the Liberal Principle of Legitimacy that "...the terms of political association must form part of a public consensus because of their essentially coercive character."[6] The project of *Political Liberalism* is an attempt to show that citizens' vastly differing conceptions of the good and the right need not lead to an indeterminacy in which principles are justified to citizens, and, therefore, eligible to govern the basic political structure of society.

The basic apparatus Rawls uses to accomplish the task of generating the necessary consensus is a distinction between the *comprehensive* and the *political*. Comprehensive doctrines contain various views about the good and the right and the details of how to live. On the other hand, the constituent principles of the common, public political conception of justice govern the

[3] Christopher McMahon (2005) has similarly argued that Pettit's single-value view leaves policy and institutional prescriptions indeterminate. However, McMahon's solution appeals to the *procedural* value of contestatory democracy. In work under preparation, I show that purely procedural values will not do the requisite work, and that a substantive value like the one defended here is required.

[4] For this sketch, I rely on the accounts of Freeman (2007), Ch.7-9, and Larmore (2003)

[5] Rawls (1993), 137. An alternate formulation is also given on p.217.

[6] Larmore (2003), 383

use of coercion. It is thus properly an object of consensus. It is legitimate only when it can be justified to reasonable citizens in terms of the constituent commitments of their individual political conceptions of justice, which are freestanding with respect to citizens' comprehensive doctrines. The Rawlsian story about legitimacy requires, in addition, that *every* reasonable comprehensive doctrine will endorse such a freestanding political conception in an "overlapping consensus". By bracketing controversial philosophical commitments about how to live, we can hopefully identify a privileged set of commitments about specifically political matters that can underwrite a consensus about the principles that govern coercive exercises of political power.

For Rawls, justification of the principles that constitute the public political conception of justice is only required in terms acceptable to those who hold *reasonable* comprehensive doctrines, where "reasonable" is a substantively normatively weighty criterion. Jonathan Quong makes the orthodox Rawlsian point in saying that *unreasonable* citizens are "excluded from the constituency of persons to whom arguments about the rights and benefits of citizenship must be justifiable."[7] Holders of unreasonable comprehensive doctrines are not owed justification in terms they can accept. Thus, the substantive normative values packed into the notion of "reasonableness" are used to separate those who are owed justification for the principles governing coercive exercise of political power by their own lights from those who are only owed justification in a second-class way that does not appeal to their commitments.[8]

The weight of the political liberal approach turns on how to delineate the bounds of "reasonableness". Which commitments, held by which people, constitute the set of acceptable justifications for the principles governing the exercise of political authority? The answers to this question are normatively loaded. For example, Rawls holds that reasonable people are those disposed to seek fair terms of social cooperation.[9] Larmore, similarly, holds that political liberalism "...rests on the principle of respect for persons, holding itself accountable therefore only to those who are committed to regulating the political use of coercion by that very principle."[10] On his view, political liberalism "...denies that the basic terms of our political life must be justifiable to citizens who reject the cardinal importance of the search for common

[7] Quong (2013); see also Quong (2011), 290-314
[8] Thanks to Dave Estlund for making me be clear about this.
[9] Rawls (1993), 49
[10] Larmore (2015), 85

ground amidst different convictions about the essence of the human good."[11] The various positions in the "public reason" literature that has erupted since the publication of *Political Liberalism* can, I think, be individuated with respect to the commitments they pack into "reasonableness".[12]

Thus, the orthodox Rawlsian view restricts the resources for justifying political principles. The public political conception of justice is justified only with respect to a truncated set of commitments--the *political* conceptions of justice endorsed by its *reasonable* citizens.

## II. Neo-Republicans Need Reasonableness

In this section, I will argue that Pettit and other neo-Republicans must appeal to an analogue to the Rawlsian notion of reasonableness. But first, let us rehearse the basic shape of Pettit's view. For the neo-Republican, the ill to be avoided by social and political institutions is *domination*. According to Pettit's specific variety thereof, one agent dominates another just in case

1. they have the capacity to interfere
2. on an arbitrary basis
3. in certain choices that the other is in a position to make.[13]

He further defines "arbitrariness" thus: "...an act of interference will be non-arbitrary to the extent that it is forced to track the interests and ideas of the person suffering the interference."[14] In particular, an interference must be "...forced to track what the interests of [the interfered-with parties] require *according to their own judgements*."[15]

[11] *Ibid.* Larmore, however, uses "reasonable" in a more expansive way than Rawls does, and does not use the term to make the distinction for which I take Rawls to use the term. Nevertheless, he *does* make the distinction. For more, see Larmore (2015), 80-86.

[12] Cf. Quong (2011); Larmore (2003) and (2015); and, for a completely different view, Gaus (2011). See also Muldoon (2012) and (2014); D'Agostino (2004) and (2005); Gaus and Vallier (2009); and Vallier (2011a), (2011b), (2012), and (forthcoming 2015); and Vallier and Thrasher (2015).

[13] Pettit (1997), 52

[14] Pettit (1997), 55

[15] *Ibid.* Emphasis mine. Presumably, if the interests to be tracked are specified according to the judgments of other people, worries about paternalism arise. Also, such a standard seems, depending on how the account is specified, far too easy to meet. By the lights of men in deeply patriarchal societies, they *are* tracking the interests of women, whom they take to be insufficiently rational, emotional creatures in need of guidance.

In order for freedom as non-domination to be the sort of thing that can serve as a standard for policy and institutions, it must regard some determinate set of interests as not worth tracking. Pettit, of course, recognizes this as a problem. He argues that such a standard must provide a language for political debate that "...employs only conceptual distinctions and inferential patterns that no one in the community has serious reason to reject."[16] More concretely, in both *Republicanism* and *On the People's Terms*, Pettit argues that the standard governing the exercise of coercive authority is justified by appeal to common interests that are not "sectional or factional in character."[17] Rather, "...acts of interference perpetrated by the state must be triggered by the shared interests of those affected under an interpretation of what those interests require that is shared, at least at the procedural level, by those affected."[18]

Pettit proposes two tests for the commonality of interests. The first is that we can simply consult the interests actually advanced by citizens in public discourse. Where there is disagreement about which of the publicly-aired interests ought to be advanced, the only recourse is, barring some sort of secession, "...a higher-level consensus about procedures."[19] Such a consensus would presumably take precedence over the base-level disagreement by providing a procedure for resolving that disagreement that advances *common* interests. The second is what Pettit calls the "tough-luck" test.[20] Throughout *On the People's Terms*, Pettit distinguishes *legitimate* coercive policy setbacks as "...not the work of a dominating will."[21] But from whose perspective is this judgment to be made? Pettit himself talks as though it ought to be made from the perspective of the party against whom the policy decision cuts. If, in relatively cool, reflective moments, the loser of a policy dispute can accept that the loss was just "tough luck," rather than the imposition of an alien will, then we, as observers, can relatively reliably conclude that the decision was the result of a more-or-less legitimate political procedure or process.

This higher-level consensus about procedures can ground an understanding of legitimacy because it plays a certain role in Pettit's theory. In particular, Pettit attributes to political actors the particular characteristic of prioritizing (to paraphrase Rorty) democracy over philosophy by holding the common interests in a particular political-institutional structure to be robustly

[16] Pettit (1997), 131
[17] *Ibid.,* 56
[18] *Ibid.*
[19] *Ibid.*
[20] Cf. Pettit (2012), 175-9, where this is laid out at most length.
[21] *Ibid.* 177

authoritative, in the political realm, over particular interests stemming from recognizably idiosyncratic views about the good. This capacity and willingness of agents to separate the set of interests that *everyone* (on Pettit's view) has from the idiosyncratic set allows the theory to appeal to the privileged, common set in order to get itself the resource of *legitimacy* in the face of disagreement about policy. As such, the higher-level consensus about procedures plays, for Pettit, a *remarkably* similar role to the one played for Rawls by an overlapping consensus of citizens' political conceptions of justice. For Rawls, political conceptions of justice just *are* the set of commonly-held values which delineate the claims owed to citizens in organizing the basic structure of society. The legitimacy of a political-institutional arrangement turns on its being organized in accordance with the relevant overlapping consensus, rather than on the broader sets of principles that constitute individuals' idiosyncratic comprehensive doctrines. The dependence of legitimacy on a particular set of individuals' interests (or principles, or values) that is the object of consensus among citizens and taken by them to have priority over their idiosyncratic counterparts is the common thread running through Pettit and Rawls.

There is, however, one crucial difference between the two. I read Pettit's explicit dichotomy between a higher-level consensus about procedures and secession to be an attempt to avoid the Rawlsian move of importing a criterion that restricts the set of interests eligible for consideration. Pettit's reliance on the single value of non-domination leaves him with more limited resources for resolving disputes about the basic structure of society. Either disputants must find a higher-level consensus about procedures that can ground a resolution, or one disputant must secede. There is a tension between this feature and his claim that non-domination provides a single political value that can be appropriately sensitive to the particular, idiosyncratic interests of cultural groups within a larger, multicultural society, *especially* those outside the cultural mainstream.[22] I am quite skeptical that even this higher-level consensus can be read off an empirical examination of the interests citizens actually advance. I suspect also that there will, in a contemporary pluralistic democracy, be plenty of disagreement about whether or not some policy setback passes the "tough luck" test. The thorny issue of the obligations of the state towards cultural minorities and other, often deeply religious (and often illiberal[23]) citizens provide a useful test case for illustrating this tension in Pettit's view.

[22] Cf. Pettit (1997), 143-146

[23] On an intuitive understanding of "illiberal".

Issues of religious accommodation often take the form of what I've called a "domination tradeoff": a choice between dominating some religious group or cultural minority and allowing some sub-group within that cultural minority to be dominated as a result of that group's internal power dynamics. Domination tradeoffs are a class of cases that illustrate the following point: If domination is the only trigger for state action, then there is nothing the state can say to justify acting in one way rather than another. Nothing about the nature of domination *per se* provides Pettit with the resources for resolving a dispute.

A nice example of this worry comes from Susan Moller Okin's "Is Multiculturalism Bad for Women?":

> "During the 1980s, the French government quietly permitted immigrant men to bring multiple wives into the country....Once reporters finally got around to interviewing the wives, they discovered...that the women affected by polygamy regarded it as an inescapable and barely tolerable institution in their African countries of origin, and an unbearable imposition in the French context. Overcrowded apartments and the lack of private space for each wife led to immense hostility, resentment, even violence both amongst the wives and against each other's children."[24]

We can extend the case slightly to show how it poses a problem for Pettit's claim that freedom as non-domination is a uniquely common political ideal that is nevertheless capable of allowing each idiosyncratic voice to be heard. Polygamy was, at that time, a group right afforded to Muslim immigrants from North Africa and to nobody else, for the purposes of preserving a cultural institution. Pettit is quite friendly to the idea that non-domination might require the provision of group rights, claiming that "...no one should baulk at the possibility that if those in certain groups are to share in the common republican good of non-domination--the common good, if you like, of citizenship--then their special positions may require that they be given special attention and support."[25] The Maghrebin community's status as an ethnic and cultural minority group within the avowedly secular French state made them vulnerable to domination. The strictures of the French constitution were such that the French government had the capacity to interfere with the cultural institution of polygamous marriage in an arbitrary fashion--that is, in a way that did not track Maghrebin cultural interests. From the perspective of Maghrebin men,

[24] Okin (1999), 9-10
[25] Pettit (1997), 145

the political value of non-domination required that, as a matter of law, the French government permit polygamous marriage. In doing so, freedom as non-domination serves as a political value that allowed Muslim immigrants to France to articulate their particular, culture-specific grievances. It was thus "...necessary...to treat members of the minority culture as conscientious objectors to mainstream ways, and to allow them various forms of exemption from otherwise universal obligations."[26]

So the toleration of polygamy was an expansion of freedom for Maghrebin immigrants--*male* Maghrebin immigrants. For the women, the toleration of polygamy represented an increase in vulnerability to domination. Polygamous marriages left women vulnerable to the whims of their husbands and *at best* in a sort of *modus vivendi* equilibrium of reciprocal power with other wives--the very sort of situation Pettit himself rejects as unsatisfactory.[27] Clearly the institution of polygamy allowed for the domination of Maghrebin women by Maghrebin men, and perhaps also by *other* Maghrebin women.

Pettit is right to note that taking up the cause of non-domination requires identifying with other members of one's vulnerability class, because "...membership in a minority culture is likely to be a badge of vulnerability to domination."[28] But, as Okin observes, "...there is considerable likelihood of tension...between feminism and a multiculturalist commitment to group rights for minority cultures."[29] Okin casts the tension as arising between the interests of cultural minorities and "...the basic liberal value of individual freedom, which entails that group rights should not trump the individual rights of its members."[30] But the point is easily translated into Pettit's conceptual framework. People--in particular, women who are also members of minority ethnic and cultural groups--often wear *multiple* badges of vulnerability to domination, and the states of affairs that secure them from domination *qua* one vulnerability class might well make them *more* vulnerable to domination in their positions as members of another vulnerability class. As Okin's case illustrates, alleviating power differentials *between* cultural groups by allowing them special rights and privileges that protect particular cultural values might well exacerbate power differentials *within* them.

[26] *Ibid.*, 146
[27] *Ibid.*, 92-95
[28] Pettit (1997), 145
[29] Okin (1999), 10
[30] *Ibid.*, 11

Another illustrative case is *Craig v. Masterpiece Cake Shop, Inc.* and the attending controversy over whether or not private businesses can refuse service for reasons of sexual orientation. The Colorado Anti-Discrimination Act prohibits so-called "places of public accommodation" from refusing service on the basis of membership in a protected class such as sexual orientation.[31] According to the law, "places of public accommodation" is defined as "any place of business engaged in any sales to the public and any place offering services, facilities, privileges, advantages, or accommodations to the public, including but not limited to any business offering wholesale or retail sales to the public."[32] As such, under the law, private businesses cannot refuse service "based in whole or in part" on the erstwhile client's homosexuality, even if the owners of those businesses hold conservative, deeply homophobic religious commitments.[33] While conservative religious groups aren't exactly cultural *minorities* in the United States, in the same way as immigrants to France from its former African colonies, they certainly have religiously-motivated views that are both idiosyncratic and deeply held. Setting aside one's own personal views about the propriety or correctness of religious objections to homosexuality,[34] religious people who are moved by religious reasons are owed the same thing, *qua* citizens, that the rest of us are owed. By Pettit's lights, that is non-domination. That is, they are owed a choice situation in which no other agent has a capacity to interfere with them on an arbitrary basis.[35] By Pettit's lights, in requiring them to serve wedding cakes to gay couples, the Colorado Court of Appeals and General Assembly have interfered with the owners of the cake shop in a way that doesn't track their interests, *as citizens and business owners whose conception of self includes deeply homophobic religious commitments*, in aiding what they view as deeply sinful conduct. Of course, in doing so, they've relieved gay and lesbian citizens of a considerable burden of interference *by homophobic business owners* that does not track their interests.

Similarly, the rhetorical firestorm of religious conservatives in reaction to *Obergefell v. Hodges* and other gains in the fight for marriage equality is, I think, evidence of this. For

[31] Colo. Rev. Stat. § 24-34-601(2)(a)
[32] Colo. Rev. Stat. § 24-34-601(1)
[33] *Craig v. Masterpiece Cakeshop, Inc.*, No. 14CA1351, 2015 WL 4760453, at ¶28 (Colo. Ct. App. 2015)
[34] I have in mind here examples like Stephen Macedo's admonition to those who feel "silenced" or "marginalized" by the idea that "...some of us believe that it is wrong to shape basic liberties on the basis of religious or metaphysical claims" to "grow up!" (Macedo (2000), 35)
[35] Pettit (1997), 52

example, Louisiana governor Bobby Jindal wrote an op-ed in the *New York Times* arguing[36] that "large corporations recently joined left-wing activists to bully elected officials into backing away from strong protections for religious liberty."[37] The language of *bullying* is of particular interest here. Jindal explicitly attributes the relevant laws to alien wills--those of courts, nefarious denizens of corporate boardrooms, and "left-wing activists." In contrast, he couches his own position as one of tolerance of diverse viewpoints. Moreover, in the wake of *Obergefell*, Jindal *explicitly* said that "[t]he Supreme Court is completely out of control, making laws on their own, and has become a public opinion poll instead of a judicial body...If we want to save some money, let's just get rid of the court".[38] None of these sounds much like passing the "tough luck" test.

The cases I've laid out have answers that seem obvious. Pettit might well reply that *of course* the proper course of action would be to disallow polygamy (perhaps with a provision for the immigration and care of superfluous wives), because the provision of special group rights is only appropriate when it lowers non-domination overall.[39] Permitting polygamy *actually* dominates Maghrebin women, whereas prohibiting it either doesn't *really* dominate Maghrebin men, or doesn't do so enough to rule out prohibiting it. Similarly, the Colorado Supreme Court was right to prevent Masterpiece Cake Shop from discriminating against the gay couple in question, and Bobby Jindal is just incorrect--it just *is* tough luck that the Supreme Court ruled against him in *Obergefell*. But the important point I want to make is that the obviousness stems from a particular set of "factional" or "sectional" interests. Their sectionality is evidenced by Pettit's own criterion--they are not shared by *all* members of the liberal, though pluralistic, societies under examination. In particular, they are roundly rejected by those with conservative religious commitments, and who have not yet heeded Macedo's injunction to "grow up!"[40] Indeed, it is *far* from an obvious empirical truth that any set of interests that is not "sectional" or "factional" is thick enough to ground a distinction between acceptable and unacceptable policy that extends over issues of religious accommodation.

The value of non-domination itself provides no resources for deciding among claims with the structure laid out above. In order to separate the legitimate claims to non-domination from

[36] On an *extremely* loose conception of "arguing," it has to be said.
[37] Cf. http://www.nytimes.com/2015/04/23/opinion/bobby-jindal-im-holding-firm-against-gay-marriage.html
[38] Cf. http://www.msnbc.com/msnbc/bobby-jindal-lets-just-get-rid-the-court
[39] My thanks to Robert Joynt and David Estlund for making me be clear about this.
[40] See FN33 above

the illegitimate ones--the ones that appeal to tracking racist, or misogynistic, or homophobic cultural interests--Pettit must import a further value. In claiming that Maghrebin men aren't *really* being dominated by a prohibition of polygamy, or that evangelical Christian store owners aren't *really* being dominated by a prohibition of discriminating against gay people, the neo-Republican smuggles in what seems like a criterion of reasonableness in the back door without acknowledging that she is doing so. Similarly, if the "tough luck" test is just supposed to be a reliable indicator of an underlying consensus about procedures, or the basic structure, or whatever, then Jindal *et al.*'s explicitly claiming imposition by an alien will makes it hard to justify positing an underlying consensus about procedural interests. Rather, Pettit must acknowledge that Jindal and his ilk fail to meet a substantive criterion of reasonableness, and, as a result, are not the sort of people who are eligible to decide which interests are required to be tracked, and, as such, what counts as the dominating or bullying conduct of an alien will.

The prohibition of polygamy and the extension of marriage rights to gay people might be policies that track the *right* interests, but this doesn't follow from any sort of consensus about interests. It does, however follow from a set of *reasonable* interests. Because Pettit explicitly claims that an interest's not being shared by some chunk of the population is sufficient to make it an inappropriate guide for policy, the presence of deep and pervasive disagreement in various relevant policy arenas creates an impasse and an imperative for a supplementary method of securing the relevant interests. This is what a criterion of reasonableness is for. Without it, the neo-Republican account's ability to license policies and institutions is held hostage to the interests of illiberal citizens. The political liberal account is not.

Bibliography


1. Colo. Rev. Stat. § 24-34-601
2. *Craig v. Masterpiece Cakeshop, Inc.*, No. 14CA1351, 2015 WL 4760453 (Colo. Ct. App. 2015)
3. D'Agostino, Fred (2004) "Pluralism and Liberalism". In Gaus, Gerald, and Chandran Kukathas, eds. *Handbook of Political Theory*. London: Sage, 239-249
4. ---------- (2005) "Legitimacy in a Pluralistic Context". In Young, Graham, and Graham Maddox, eds. *Legitimation and the State*. Armidale: Kardooair Press, 15-29
5. Estlund, David (2014) "Review of *On the People's Terms*". *Australaisan Journal of Philosophy* 92.4, 799-802
6. Freeman, Samuel, ed. (2003) *The Cambridge Companion to Rawls*. Cambridge: Cambridge UP
7. --------- (2007) *Rawls*. New York: Routledge
8. Gaus, Gerald (2011) *The Order of Public Reason*. Cambridge: Cambridge UP
9. -------- (2015) "Public Reason Liberalism". In Steven Wall, ed. *The Cambridge Companion to Liberalism*. Cambridge: Cambridge UP, 112-140
10. Gaus, Gerald, and Kevin Vallier (2009) "The Roles of Religious Conviction in a Publicly Justified Polity: The Implications of Convergence, Asymmetry, and Political Institutions". *Philosophy and Social Criticism* 35.1, 51-76
11. Hayek, F.A. (1960) *The Constitution of Liberty*. Chicago: Univ. of Chicago Press
12. Howard, Adam (2015) "Bobby Jindal: 'Let's just get rid of the court'". *MSNBC*. Available at <http://www.msnbc.com/msnbc/bobby-jindal-lets-just-get-rid-the-court>
13. Jindal, Bobby (2015) "Bobby Jindal: I'm Holding Firm Against Gay Marriage". *New York Times* op-ed. Available at <http://www.nytimes.com/2015/04/23/opinion/bobby-jindal-im-holding-firm-against-gay-marriage.html?_r=0>
14. Larmore, Charles (2003) "Public Reason". In Freeman, ed. (2003), 368-393
15. -------- (2008) "The Meanings of Political Freedom". In *The Autonomy of Morality*. New York: Cambridge UP, 196-219
16. -------- (2015) "Political Liberalism: Its Motivations and Goals". *Oxford Studies in Political Philosophy* 1, 63-88

17. McMahon, Christopher (2005) "The Indeterminacy of Republican Policy". *Philosophy & Public Affairs* 33.1, 67-93
18. Macedo, Stephen (2000) "In Defense of Liberal Public Reason: Are Slavery and Abortion Hard Cases?". In George, Robert P., and Christopher Wolfe, eds. *Natural Law and Public Reason*. Washington D.C.: Georgetown UP, 11-49
19. Muldoon, Ryan, Chiara Lisciandra, Jan Sprenger, Carlo Martini, Giacomo Sillari and Mark Colyvan (2014) "Disagreement Behind the Veil of Ignorance". *Philosophical Studies* 170.3, 377-394
20. Mudoon, Ryan, Michael Borgida, and Michael Cuffaro (2012) "The Conditions of Tolerance". *Philosophy, Politics, and Economics* 11.3, 322-344
21. *Obergefell v. Hodges*, 576 U.S._____ (2015)
22. Okin, Susan Moller (1999) *Is Multiculturalism Bad For Women?* Princeton: Princeton UP
23. Pettit, Philip (1997) *Republicanism*. Oxford: Oxford UP
24. -------- (2006) "The Determinacy of Republican Policy: Reply to McMahon". *Philosophy & Public Affairs* 34.3, 275-283
25. -------- (2012) *On the People's Terms*. Cambridge: Cambridge UP
26. Pettit, Philip, and Christian List (2011) *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford: Oxford UP
27. Quong, Jonathan (2011) *Liberalism Without Perfection*. Oxford: Oxford UP
28. -------- (2013) "Public Reason". In Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy*
29. Rawls, John (1971) *A Theory of Justice*. Cambridge, MA: Harvard UP
30. -------- (1980) "Kantian Constructivism in Moral Theory". *Journal of Philosophy* 77.9, 515-572
31. -------- (1993) *Political Liberalism*. New York: Columbia UP
32. -------- (1997) "The Idea of Public Reason Revisited". *University of Chicago Law Review* 64.3, 765-807
33. Rorty, Richard (1988) "The Priority of Democracy to Philosophy". In Peterson, Merill, and Robert Vaughan, eds. *The Virginia Statute of Religious Freedom*. Cambridge: Cambridge UP, 257-288

34. Skinner, Quentin (1998) *Liberty Before Liberalism*. Cambridge: Cambridge UP
35. Vallier, Kevin (2011a) "Against Public Reason's Accessibility Requirement". *Journal of Moral Philosophy* 8.3, 366-389
36. -------- (2011b) "Consensus and Convergence in Public Reason". *Public Affairs Quarterly* 25.4 (Symposium on Convergence Justifications in Public Reason), 261-279
37. -------- (2012) "Liberalism, Religion, and Integrity". *Australasion Journal of Philosophy* 90.1, 149-165
38. -------- (2016) "In Defense of the Asymmetric Convergence Model of Public Justification: Reply to Boettcher". *Ethical Theory and Moral Practice* 19.1, 255-266
39. Vallier, Kevin, and John J Thrasher (2013) "The Fragility of Consensus: Public Reason, Stability, and Diversity". *European Journal of Philosophy* 21.2, online
40. Weithman, Paul (2010) *Why Political Liberalism?* Oxford: Oxford UP

Coleen Macnamara
UC Riverside

**Bio:** Coleen Macnamara is an Associate Professor of Philosophy at the University of California, Riverside. She works in moral philosophy. Her research focuses on the pluralistic nature of reasons, Strawsonian reactive attitudes, and the nature of our practice of holding others responsible.

# Hypocrisy and the Standing to Blame

## §1

Moral agents are bound by norms. But this is not the entirety of our moral system. We are also accountable to each other for complying with those norms: members of the moral community have standing to blame norm-violators. This is, as Stephen Darwall puts it, "morality as equal accountability" (2006, 97). Moral agents are bound not only *by* norms but also *to* each other.

While moral agents have default standing to blame others for their moral wrongdoings, this standing is not inalienable. Imagine Darlene lies to Mia and refuses to repent. Subsequently, Mia lies to Darlene and Darlene swells with resentment, and lets Mia know it. Mia will likely react to Darlene's blame not with guilt and apology, but with "Who do you think you are? You have no right!" On the received view, Mia is correct. Darlene is a hypocrite. In virtue of her unrepentant wrongdoing, Darlene loses the standing to resent Mia. Not only is Darlene's *expression* of resentment untoward; the resentment itself is inappropriate. This is not to say that Mia has done nothing wrong. She certainly oughtn't lie to Darlene. The point is that while others may have standing to blame Mia, *Darlene* does not.

We might wonder why unrepentant wrongdoers lose the standing to –even privately – blame those who commit parallel wrongs? In this paper, I identify a key element of an answer to this question. I urge that a core function of morality as equal accountability is the constitution of a social reality that conduces to the development of

fully self-respecting moral agents. An unrepentant wrongdoer loses her standing because her retaining said standing thwarts this core function. [1]

## §2

As is so often the case in philosophy, making progress will require backing up and slowing down. Our starting point is the intuition that unrepentant wrongdoers lose something we all presumptively have. Let's take a moment to articulate precisely just what this something is.

To begin, we can say that they have lost the standing to blame. We can say more about both blame and standing. In this paper I adopt the widely endorsed reactive attitudes account of blame, according to which blame is, in the first instance, the emotional response of resentment or indignation. We blame both when we express our resentment ("You jerk!") or indignation ("How dare you!"), and when we keep them buried in our hearts.[2]

Standing is a positional notion. The standing to blame is a position, relative to a particular blameworthy agent, that one occupies in virtue of possessing a kind of authority. The idea is familiar in the realm of punishment. Regular Joe Citizen does not have the standing to punish criminals no matter how punishment-worthy they may be. Only those who possess the requisite authority – duly appointed agents of the state – are in a position to impose punishments.[3] Just so, only those with the relevant authority have the *standing* to blame.[4]

[1] For competing accounts see Wallace (2010) and Scanlon (2008). In a longer version of this paper, I urge that these accounts are inadequate.

[2] The blaming response appropriate to victims of wrongdoing is resentment. Indignation is the response appropriate to third parties. In this paper, I exclusively focus on why unrepentant wrongdoers lose the standing to *resent* as opposed to feel *indignation.* In other words, I focus on why an unrepentant wrongdoer loses the standing to resent her victims for committing parallel wrongs toward her. Though I have not worked out all the details, I am confident my view can be extended to the standing to blame via indignation.

[3] Indeed, we have a pejorative term for those who exact punishment without the requisite authority: we call them vigilantes.

[4] Standing then is different from other factors that affect the appropriateness of a particular instance of blame. The blameworthiness of the blamee, another necessary condition of appropriate blame, is not about authority. Blameworthiness depends instead on the one hand on whether the potential blamee is a morally responsible agent and responsible for the particular piece of conduct in question. And on the other hand, on whether said conduct carries the right kind of *moral significance.* If the potential blamee is child, or again manipulated by an

Already we can say more precisely what the unrepentant wrongdoer loses: she loses a particular kind of authority. [5] However, I think we can identify the authority lost more precisely as the authority defining of the accountability-to relation. To say 'A has standing to blame B for φ -ing' is to say that B is accountable to A for φ -ing. The accountability-to relation is an authority relation. To say that B is accountable to A is to say that A has a kind authority relative to B. It is this authority that grounds A's standing to blame. Losing this authority means losing the standing to blame.

The accountability-to relation, though, involves not only blame, but also just as crucially *apology*. There are two relata in the accountability-to relation: the "accounter" and the "accountee." Familiarly, when the accountee wrongs the accounter, she ought to apologize. Equally familiarly, it is not just that she *ought* to apologize, but that she *owes* the accounter an apology. Here "owing" marks the fact that the ought in question is second-personal or relational in the sense that it is grounded in the accounter's authority. It is, then, the authority defining of the accountability-to relation that grounds the second-personal ought to apologize.

This authority is crucial to another familiar component of our accountability practices: normative expectations**.** The term 'normative expectation' is often used to refer to the requirements of morality or to moral oughts generally. However, that is not how I am using it here. Instead, I am using it to refer to what Wallace (1996) calls the "stance of holding another responsible." On this use, normative expectation is a distinctive psychological attitude. It is a forward-looking way of holding another to the ought that binds her, just as the reactive attitudes are backward-looking ways of doing the same thing. The best way to unpack this is to look at examples in which normative expectations are abandoned for prudential reasons.

evil scientist, or yet again not if the conduct is not morally untoward, then the potential blamee is not blameworthy. None of this is about authority relation between one person and another.

The appropriateness of a particular instance of blame is also beholden to the normative valence of the exercise of authority that blaming constitutes. Just as it is sometimes morally or again prudentially untoward to exercise a right one has – think of refusing move to another seat on a train so that an elderly couple can sit together – so too it can be morally or prudentially untoward to exercise the authority one has to blame. But as with blameworthiness this issue is not, as standing is, about whether or not one has the requisite authority vis-a vis to another. It is rather about the appropriateness of exercising the authority we all agree one has.

[5] In what follows I paint a picture of what this authority is. We might also say what it is not. I have argued elsewhere that contra Darwall, this authority is not the authority to issue demands. See for example, Macnamara (2013a and 2013b).

Consider Alice and Bianca. Alice is a serial liar. She lies frequently about small matters and occasionally about weighty ones. Bianca know this about Alice, and though she has tried to get Alice to change her lying ways, nothing has worked. Bianca is at her wits' end: fed up with constantly feeling resentment and the testy interactions that ensue when she expresses that resentment. Bianca wants serenity and happiness in her life, not resentment and quarrels.

Certainly Bianca could solve her problem by cutting off relations with Alice. But supposing that she is for whatever reason is unwilling or unable to do so, there is another option open to her. Bianca vows to keep her resentment to herself – effectively instituting a no-scolding policy. Though this is ends the testy interactions, another problem remains: Bianca is still subject to the deeply unpleasant feelings of resentment and still finds herself forced to suppress the consequent urge to engage in the sort of conflicts she foreswore. In order to prevent resentment from ever rearing its ugly head, she decides to emotionally disengage, to "let it go." More technically, she abandons the attitude of normatively expecting that Alice not lie.

To be clear, Bianca's *predictive* expectations have not changed. She retains a strongly grounded predictive expectation that Alice will lie (Wallace 1996, 20-21). Nor has she revised her beliefs about the relevant norm. She continues to believe that lying is just as wrong for Alice as anyone else. Rather, Bianca is letting go of the attitude that leaves her susceptible to feeling resentment (Wallace 1996, 18–40). She stops normatively expecting that Alice not lie.

Though we often have latitude regarding whether or not to abandon a normative expectation, aptly adopting this attitude requires authority. This is because normative expectations are a distinctive way of being emotionally invested in someone doing as she ought. They have a unique emotional profile. When someone violates a normative expectation we don't just feel sad, we resent her. To normatively expect that another do as she ought is to adopt an attitude toward another that gestalts the relevant action as one that is *owed* to oneself. The owing relation is an authority relation: the owed has a kind of authority with respect to the owee. Thus, to normatively expect is to assume a position that requires a kind of authority relative to the normative expectee. It is only apt to assume this position when one in fact possesses said authority.

The authority at issue here is, once again, the authority defining of the accountability-to relationship. Normative expectations are a core part of our accountability practices, conceptually connected to and on par with blaming. Just as the authority that grounds the standing to blame is the authority defining of the accountability-to relationship, so too with normative expectations.

We started with the question "What do unrepentant wrongdoers lose?" Our preliminary answer was that they lose the standing to blame. Reflection on the nature of standing brought us to the idea that the unrepentant wrongdoer loses a kind of authority. This authority, we found, is the authority defining of the accountability–to relation. The considered answer to our question, then, is that the unrepentant wrongdoer loses this authority, an authority that grounds the standing to blame, the standing to normatively expect, and the relational ought to apologize that binds the wrongdoer.

This claim is supported by returning once again to Mia and Darlene. Darlene lies to Mia and refuses to repent. In the introduction we imagined that Mia subsequently lies to Darlene and Darlene responds with resentment and its expression. We explained that Mia would likely react to Darlene's response with "Who do you think you are? You have no right!" In other words, Mia would respond by telling Darlene that she lacks the standing to blame her.

Imagine now a revised case. Keep everything the same except that this time Darlene is too emotionally tired to feel resentment. What she is not too tired to do is say to Mia, "You owe me an apology." In all likelihood, Mia is going to balk, saying, "Have you forgotten that you lie to me all the time without even feeling the tiniest tinge of guilt? And I am still waiting for *your* apology." Mia, in other words, will respond by telling Darlene that she does not owe her an apology. It certainly seems that Mia is right. To be sure, there might be other considerations that make it the case that Mia ought to apologize – keeping one's own side of the street clean comes to mind. This, though, is beside the point. It is not that apology is not required, but rather that it is not *owed*. Unrepentant wrongdoing has undermined the authority grounding the *second-personal* or *relational* ought to apologize.

Finally, imagine that before Mia even has a chance to lie to Darlene, circumstances arise that prompt Darlene to say "Mia, I expect you to always be truthful."

Again, in all likelihood, Mia will have none of this. She will respond, "Have you lost your mind? Think about all the lies that come out of your mouth." Mia, in other words, will likely respond by telling Darlene that Darlene lacks the standing to normatively expect that Mia not lie.

Over the course of this vignette, Mia has told Darlene that she lacks the standing to blame, the standing to normatively expect, and that she, Mia, no longer owes Darlene an apology. Mia has, in other words, told Darlene that her unrepentant wrongdoing has stripped her of the authority defining of the accountability-to relation. If Mia is right – and it certainly seems that she is – then our considered answer to "What do unrepentant wrongdoers lose?" is correct.

## §3

It is now apparent that we can more aptly reframe the animating question of this paper as "Why do unrepentant wrongdoers lose the authority defining of the accountability-to relation?" Over the course of the next two sections, I propose an answer to this question. The proposal, though, is arrived at indirectly. In trying to understand why unrepentant wrongdoers lose this authority, I think we will be well served to first get a handle on why moral agents presumptively possess it. So, let's start there: Why is the authority defining of the accountability-to relation theirs to lose in the first place?

To ask this question is to ask why our moral system is constituted not only by the norms that bind us, but also by morality as equal accountability. To say that our moral system is one of "equal accountability" just is to say that moral agents presumptively occupy the position of accounter with respect to other members of the moral community

One might invoke intrinsic worth to explain why we each presumptively occupy the position of accounter. On this view, our intrinsic worth grounds not only first-order moral norms, but also the accountability-to relation. We each have the kind of authority defining of the accountability-to relation because this is what our particular sort of value entails. The core assumption of this paper – namely, that unrepentant wrongdoers no longer occupy the normative position of accounter – speaks against this view. First, the intrinsic worth of unrepentant wrongdoers remains intact. It is thus not

clear why unrepentant wrongdoing would undermine one's normative position relative others. This, though, is not the view's only difficulty. Return to Mia and Darlene. Darlene lies to Mia and refuses to repent. Consequently, Darlene no longer occupies the accounter position with respect to Mia and the requirement not to lie. This, though, does not mean that Mia is now free to lie to Darlene: Mia is still bound by moral norms. Crucially, though, on the view we are considering, Darlene's intrinsic worth grounds both the requirement Mia faces *and* Darlene's authority with respect to Mia. If this is true, it is hard to see how it could be that Mia is still bound while Darlene's authority is undermined. If both the requirement and the authority have the same grounding, then it seems they should rise and fall together.

Let's then shift gears and try what, I think, is a far more promising route. Perhaps morality as equal accountability, like our practice of promising, is a social practice. If this is correct, then morality as equal accountability, like our social practices more generally, earns its normative keep via the functions it performs. On this view, while our moral norms may be grounded in our intrinsic value, the presence of morality as equal accountability is traced to the value it promotes.

Theorists have said several things about the value of our accountability practices. Wallace (1996), Franklin (2013), and Bell (2013), for example, emphasize the ways in which these practices allow us to give voice to, articulate, affirm, protect and even deepen our commitment to our moral values. Walker (2006), Scanlon (2008), and Calhoun (1989) draw attention to the ways these practices catalyze moral repair. Our interest here, though, will be best served by bringing to light and focusing on a different function of morality as equal accountability: the function of constituting a social reality that conduces to the development of fully self-respecting moral agents.

It is a fundamental tenet of morality that you and I, qua rational agents, have equal intrinsic moral worth. To be a self-respecting moral agent is, in part, to have uptake of one's equal moral worth. Robin Dillon, however, has urged that such uptake involves more than simply coming to believe that one has equal worth. It involves more than *intellectual* understanding. Fully self-respecting agents have *experiential* understanding of their equal worth. They do not just *know* that they possess equal worth, they *feel* it (Dillon 1997, 227). Just as it is one thing to know that a loved one has died and yet

another to experience the death and its import, so too it is one thing to know that you have equal intrinsic worth and another to fully experience it.[6] Experiential understanding of our equal worth goes beyond mere doxastic comprehension. It is a way of being in the world that involves, among other things, interpretive, emotional and motivational elements.

Because we are embodied, social beings, experiential understanding of equal intrinsic worth requires that our social reality reflect this status. We can experience our full self-worth only if we live it, and we can live it only if the structures and workings of the social space are such that it is enacted.

Morality as equal accountability is one of the structures within social space that encourages the creation of a social reality that conduces to the development of full self-respect.[7] Recall that the authority defining of the accountability-to relation grounds the standing both to normatively expect that others do as they ought and to resent them when they flout this ought. It is widely recognized that the latter is a way of experientially inhabiting one's intrinsic worth.[8] To resent another is, among other things, to first-personally inhabit one's intrinsic worth: it is a way of emotionally enacting one's self-respect. To borrow Joel Feinberg's vivid phrase, when we resent, "we are standing up like men," protesting another's maltreatment of us in a way that insists upon our worth (1970, 252). Similarly, when we normatively expect that another treat us as she ought, we first-personally inhabit our intrinsic worth. In adopting this attitude we live our conviction that there are boundaries that others may not cross.

Normative expectations and resentment are also ways of inhabiting the normative position at the heart of morality as equal accountability. Recall that to normatively expect is to adopt an attitude toward another that gestalts the relevant action as one that is *owed* to oneself. To adopt this attitude is to inhabit one's position of authority relative to the normative expectee. Similarly with resentment. It is widely recognized that resentment has a "call-and-response" structure.[9] It *calls* on its target – the wrongdoer – to respond

[6] I borrow this example from Dillon (1997).
[7] This idea can be found in others' work. See, for example Honneth (1992) who builds on Hegel's work.
[8] Add cites.
[9] See, for example Walker (2006, 135), Darwall (2006, 159), McGeer (2012, 303), Smith (2008, 81), and Kukla and Lance (2009).

with apology.[10] This structure brings into focus the fact that resentment is a form of first-personal practical uptake of one's normative position as an accounter. It is because we occupy this position that it is apt for us to *call on* the wrongdoer for apology.

Sincere apology is, in the first instance, a way of acknowledging one's wrongdoing to the wronged. In doing this, one is doing quite a lot. First, in acknowledging wrongdoing, one recognizes the other as having intrinsic worth. Her intrinsic worth rendered the conduct wrong. Second, if the sincere apology is issued because it is owed, or again, as an RSVP to the wronged agent's resentment, then in apologizing one is recognizing the wronged as an accounter, and first-personally inhabiting one's own position as accountee.[11]

This is not mere *receptive* recognition of the wronged agent's intrinsic worth and position. Apologies are a form of discursive, second-personal recognition. They not only recognize the wronged agent as having intrinsic worth and occupying the position of accounter, but also reflect that recognition back to the wronged agent. Consider the difference between receptively recognizing me as Coleen as we pass in the hall and second-personally recognizing me as Coleen when you say, "Hey, Coleen" The latter constitutes a kind of second-personal recognition in which you discursively convey, or reflect, your recognition of me *to* me. Apologies constitute a kind of second-personal recognition in which the wrongdoer conveys, or reflects, her recognition of the wronged agent's worth and position *to* the wrongdoer.[12]

The second-personal recognition constitutive of apology plays a particularly important role in the construction and solidification of our practical self-relations. I come to see myself as Coleen largely because others have second-personally recognized me as Coleen. Similarly, it is largely through the (ongoing, repeated) social practice in which others second-personally recognize us as individuals with intrinsic worth and the relevant normative positions, that we come to recognize ourselves as such (Kukla and Lance 2009, 178-195).

Our accountability practices, then, consist in first and second-personal uptake and recognition of our own and others' intrinsic worth and position. Crucially, though,

[10] See Smith (2007), Walker (2006), Shoemaker (2007), Darwall (2006), and Macnamara (2013a).
[11] I follow Darwall in referring to the response as an RSVP (2006, 40).
12 For more on discursive recognition, see Kukla and Lance (2009, ch. 6).

morality as equal accountability consists not just in isolated moments of uptake and recognition. Rather, these are elements of a self-reinforcing structure.[13] To normatively expect that another φ is to first-personally inhabit one's worth and position. Normative expectations, moreover, leave one susceptible to resentment, which is itself a way of inhabiting one's worth and position. Resentment in turn specifically solicits an apology from its target. Apology, when given because it is owed or as an RSVP to resentment, second-personally recognizes the worth and position of the wronged. Apology thus serves to reinforce the wronged agent's conception of herself as having the very status and position that normative expectations embody. Normative expectations lead to resentment; resentment leads to apology; and apology leads back to normative expectations. This self-reinforcing structure makes it far more likely that morality as equal accountability will create the practical self-relations at which it aims.

Morality as equal accountability also conduces to a social reality that renders our *equality* phenomenologically vivid. Proper self-respect is relational: it involves not just properly appreciating one's own status and position, but also fully appreciating that others are one's equals. If one experiences one's own status and position, but also experiences oneself as better/lesser than, or again, as dominant/subordinant to others, then one's experience does not reflect moral reality.

Morality as equal accountability conduces toward experiences of our *own* intrinsic worth because it encourages us to first-personally inhabit our worth and requires others to second-personally recognize said worth. It also conduces towards a lived experience of *others*' worth – for example, when we apologize to others, thereby second-personally recognizing their intrinsic worth. These two forms of experience can lead to an understanding of equal worth. I experience my intrinsic worth and that of others, realize that we are equally intrinsically valuable, and thus come to know that we have equal status.

Morality as equal accountability also conduces to experiences of ourselves as each occupying a *position* equal to that of our fellows, i.e. the position of accounter. When we mutually first-personally inhabit our position and second-personally recognize others as occupying this position, we experience our equality. This experience, though,

13 Thanks to Joshua Hollowell for helping me to see the importance of this point.

has elements that distinguish it from our experience of equal intrinsic worth. Unlike intrinsic worth, the position of accounter is *inherently relational.* My experience of my intrinsic worth is not, in and of itself, an experience of my relation to you. In contrast, when I inhabit my position as accounter, I inhabit a position relative to you and thus experience myself in relation to you. I experience myself as an accounter only inasmuch as I experience you as an accountee. In isolation from other experiences this is not an experience of equality, but of *inequality*. The accounter/accountee relationship is inherently unequal. It is akin to a creditor-debtor relation insofar as the accountee *owes* it to the accounter to do as she ought and *owes* the accounter an apology if she does not.

What transforms this experience of inequality into one of equality is precisely the social reality that morality as *equal* accountability aims to create. Morality as equal accountability accords to all, *ceteris paribus,* the position of accounter. In doing so, it conduces toward a reality in which a moral agent experiences not only herself as accounter and you as an accountee, but also you as accounter and herself as accountee. Morality as equal accountability is structured to create a world in which we *mutually* first-personally inhabit our positions and second-personally recognize our fellows as occupying these positions. When this happens, we live our equality. This experience of equality is, moreover, an especially vivid one because our equality has been instantiated in the social world in an equal relationship.

## §4

The above has urged that our intrinsic worth, though it may serve to ground our moral norms, does not suffice to explain morality as equal accountability. Morality as equal accountability is, I suggested, best understood as a social practice that, like other social practices, earns its normative keep via its functions. I next argued that an often-overlooked but core function of morality as equal accountability is the creation of a social reality that conduces to the development of fully self-respecting moral agents. With this much in hand, we are set (at last) to directly address the animating question of the paper, "Why do unrepentant wrongdoers lose the position of accounter?"

Imagine a world in which unrepentant wrongdoers retain the position of accounter. Darlene lies to Mia, feels no guilt, and refuses to respond to Mia's resentment with apology. Consequently, Mia lies to Darlene. Now, imagine that Darlene *retains* the standing to resent Mia and that Mia owes Darlene an apology. In this world, Darlene both refuses to first-personally inhabit her position as accountee and to second-personally recognize Mia's intrinsic worth and accounter position. Nonetheless, in our scenario, she has license both to recognize Mia as accountee and to first-personally inhabit her own intrinsic worth and accounter position. Mia, moreover, is required both to first-personally inhabit her position as accountee and to second-personally recognize Darlene's intrinsic worth and accounter position.

Such a one-sided state of affairs conduces toward a lived experience of *inequality* – one that runs quite deep. If Darlene acts as she is licensed to do, and Mia does as required, then Darlene will experience herself as having greater intrinsic worth than Mia and Mia will experience herself as having less intrinsic worth than Darlene. These are paradigmatic experiences of unequal status.

Mia and Darlene will experience themselves not only as persons of unequal status, but also as occupying unequal positions. To appreciate the full import of this, recall that the normative position of accounter is intrinsically relational. When there is an accounter, there is also an accountee. The accounter/accountee relation is, moreover, inherently unequal. The accounter has a kind of *power over* the accountee. Thus, if Darlene does as she is licensed to do and Mia as she is required to do, Darlene will experience herself as dominant and Mia herself as subordinate.

A social structure that allows Darlene, the unrepentant wrongdoer, to retain her position, conduces toward the creation of a social reality in which there are lived experiences of inequality. Such reality allows, unrepentant wrongdoers, like Darlene, to experience themselves as having a higher status and dominant position. And allows their victims, like Mia, to experience themselves as having a lesser status and subordinate position.

If allowing unrepentant wrongdoers to retain their position carries this result, it is clear why morality as equal accountability allows no such thing. I have urged that one of the core points of the practice is to create a social reality that conduces to the

development of fully self-respecting moral agents. It therefore makes sense that unrepentant wrongdoers lose their standing, because allowing them to retain it would be deeply antithetical to the goal of the practice.

**Works Cited**

Bell, M. 2013. The Standing to Blame: A Critique. In D.J. Coates and N. Tognazzini (eds.), *Blame: Its Nature and Norms* (pp. 263– 281). New York, NY: Oxford University Press.

Calhoun, C. 1989. Responsibility and Reproach. *Ethics* 99: 389-406.

Darwall, S. 2006. *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, MA: Harvard University Press.

Dillon, R.S. 1997. Self-Respect: Moral, Emotional, and Political. *Ethics* 107(2): 226-249.

Feinberg, J. 1970. The Nature and Value of Rights. *Journal of Value Inquiry* 4(4): 243-257.

Franklin, C. 2013. Valuing Blame. In D.J. Coates and N. Tognazzini (eds.), *Blame: Its Nature and Norms* (pp. 207-223). New York, NY: Oxford University Press.

Honneth, A. 1992. *The Struggle for Recognition: The Moral Grammar of Social Conflicts*. Cambridge, MA: MIT Press.

Kukla, R. and Lance, M. 2009. '*Yo!' and 'Lo!': The Pragmatic Topography of the Space of Reasons*. Cambridge, MA: Harvard University Press.

Macnamara, C. 2013a. 'Screw You!' & 'Thank you.' *Philosophical Studies* 165(3): 893–914.

_____ 2013b. Taking Demands Out of Blame. In D.J. Coates and N. Tognazzini (eds.), *Blame: Its Nature and Norms* (pp. 141– 161). New York, NY: Oxford University Press.

McGeer, V. 2012. Co-reactive Attitudes and the Making of Moral Community. In C. MacKenzie & R. Langdon (eds.). *Emotions, Imagination and Moral Reasoning* (pp. 299–326). Macquarie monographs in cognitive science. Psychology Press.

Scanlon, T. M. 2008. *Moral Dimensions: Permissibility, Meaning, and Blame*. Cambridge, MA: Harvard University Press.

_____2013. Interpreting Blame. In D.J. Coates and N. Tognazzini (eds.), *Blame: Its Nature and Norms* (pp. 141– 161). New York, NY: Oxford University Press.

Smith, A. 2007. On Being and Holding Responsible. *The Journal of Ethics* 11:265-484.

_____ 2008. Control, Responsibility, and Moral Assessment. *Philosophical Studies* 138: 367–392.

Shoemaker, D. 2007. Moral Address, Moral Responsibility, and the Boundaries of the Moral Community. *Ethics* 118(1): 70–108.

Wallace, R. J. 1996. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

_____ 2010. Hypocrisy, Moral Address, and the Equal Standing of Persons. *Philosophy and Public Affair*s 38(4): 307– 341.

Walker, M. U. 2006. *Moral Repair: Reconstructing Moral Relations After Wrongdoing*. Cambridge: Cambridge University Press.

David Beglin
UC Riverside

**Bio:** David Beglin is a PhD candidate at University of California, Riverside, where he is currently writing a dissertation on the nature of blame and morally responsible agency. In his dissertation, David is developing an account of these things by drawing on the work of P.F. Strawson, including both Strawson's well known "Freedom and Resentment" but also Strawson's metaphilosophical works. More broadly, David's research interests include ethics, moral psychology, agency theory, and the philosophy of death.

# The Nature of Blame and Our Reasons for Forgiveness

## 1. Introduction

Theorists have long been interested in what is required for someone to be the type of agent whom it is appropriate to blame for her behavior. Recently, many of these theorists have looked to blame's nature to provide an account of these requirements. The idea is intuitive. If we want to know what makes someone an appropriate object of blame, we should first get clear on what blame is—particularly on what is at stake in our blame.[1] One of the most prominent instances of this strategy is what I'll call the "Communication View."[2] According to this view, blame is a form of moral address, communicating a message to the wrongdoer, and it is from blame's communicative nature that the standards for being an appropriate object of blame arise.

While I don't doubt that blame has some communicative element to it, I nevertheless worry that the Communication View overstates the importance of this communicative element to our blaming practices. To motivate this worry, I'll look to forgiveness. The nature of blame appears to be conceptually connected to the nature of forgiveness. And the particular way these phenomena are conceptually connected, I

[1] Throughout this paper I use phrases like "appropriate object of blame" or "blame's standards of propriety." These phrases are ambiguous. One might be an appropriate object of blame on some particular occasion, perhaps because one performed some wrongful action, but one might also be an appropriate object of blame in a more general sense—one might be a morally responsible agent, the kind of agent whom would be appropriate to blame on some particular occasion *if* she performed a wrongful action. I'm concerned with this second sense of "propriety" or "appropriateness."

[2] For proponents of this sort of view, see: Watson 1987 and 2011; Darwall 2006; Shoemaker 2007; McKenna 2012; Fricker 2016; Macnamara 2015a.

argue, puts a condition on any theory of blame's nature: any theory of the nature of blame must elucidate our reasons for forgiveness. Blame's communicative element, though, I argue, doesn't feature in our reasons for forgiveness. And this suggests it may not be as central to blame's nature as the Communication View takes it to be.

I thus have two aims in this paper. First, I hope to raise a worry about the Communication View. But second, and arguably more importantly, I hope to make a positive point about the connection between forgiveness and blame, a point about how these two phenomena are conceptually connected and about how this puts constraints on our theorizing about them.

**2. The Communication View**

I turn to the conceptual connection between blame and forgiveness in the next section. First, we should get a better sense of the Communication View and its commitments.

Gary Watson first introduced the Communication View as a way of filling a purported gap in the account of blame P.F. Strawson put forth in his pivotal lecture, "Freedom and Resentment." In that lecture, Strawson holds that blame paradigmatically takes the form of certain emotions, "reactive attitudes," such as resentment or indignation. These reactive attitudes reflect a demand for goodwill or regard, according to Strawson, and he suggests that those whom we exempt from our blaming practices, those people we don't consider appropriate objects of blame, even when their actions fail to show us due regard, are those who are excepted from this demand.[3] Watson argues that Strawson's theory is incomplete. Particularly, he suggests that Strawson fails to explain the conditions for exemption: Strawson tells us who typical exempt agents are—children, for instance, or people suffering from various mental illnesses—and he tells us that these agents are exempt from blame because they aren't subject to the demand for regard or goodwill the reactive attitudes reflect; but he doesn't tell us *why* those agents aren't subject to the demand and thus exempt from blame. Strawson, according to Watson, doesn't tell us "what kind of explanations exempt" or "how this works" (Watson 1987,

[3] Strawson's discussion of these issues is notoriously difficult. Here, I follow Watson in my understanding of Strawson. There is certainly more to say, though. For passages relevant to the reactive attitudes and the way they relate to the demand for goodwill or regard, see Strawson 1962, 48-50, 56-58, 63.

228).[4] To fill this lacuna in Strawson's theory, Watson suggests we "construe the exempting conditions as indications of the constraints on intelligible moral demand, or, put another way, of the constraints on moral address" (229). Watson, in other words, introduces the Communication View.

Watson, extending Strawson's thought, suggests that the reactive attitudes are "forms of communication, which make sense only on the assumption that the other can comprehend the message" (Watson 1987, 230). It is not, then, just that the reactive attitudes reflect a demand. Rather, Watson takes the reactive attitudes to aim at communicating that demand to their object. And demanding in this way, Watson notes, "presumes understanding on the part of the object of the demand" (230). Intelligibly communicating a demand to another person, in other words, requires that the other person be capable of understanding what is being demanded of them. If someone is incapable of comprehending the demand for regard the reactive attitudes communicate, then that person is an inappropriate object of that demand and, thus, an inappropriate object of the reactive attitudes. In such a case, Watson explains, the reactive attitudes "lose their point as forms of moral address" (231).

Many theorists have followed Watson in adopting the Communication View, and there are interesting differences between all of them.[5] Here, however, I'm concerned with what they have in common. Underlying all of these theorists' accounts are two claims.[6] I

[4] There is some reason to doubt Watson's worry about Strawson's view. For instance, Strawson seems to take the capacity to participate in ordinary relationships to be a condition for blame's propriety: "Now it is certainly true that in the case of the abnormal, though not in the case of the normal, our adoption of the objective attitude is a consequence of our viewing the agent as *incapacitated* in some or all respects for ordinary inter-personal relationships" (Strawson 1962, 55). (Here, we can understand exemption in terms of adopting the objective attitude (cf. Strawson 1962, 50-53).) Nevertheless, for my purposes, whether Watson's worry about Strawson is correct isn't relevant.

[5] Stephen Darwall, for instance, argues that the reactive attitudes are "quasi-speech acts" that "lose their point" if one lacks "second-personal competence" (Darwall 2006, 75-76). Similarly, David Shoemaker suggests that the reactive attitudes are "pointless as a form of moral address" if their object lacks the capacity for a certain kind of empathy and to be motivated by this empathy (Shoemaker 2007; 75; 107). And recently, Coleen Macnamara has argued that one is "ineligible for the role of blamee" if one is "ineligible for the role of addressee," and that one is only eligible for the role of addressee if one is capable of giving uptake to the response the reactive attitudes invariably call for—guilt (Macnamara 2015a, 212).

[6] Here, I'm bracketing another commonality between all of these theorists. They all take blame to paradigmatically involve the reactive attitudes. For my purposes, this shared commitment isn't relevant.

take these claims to be definitive of the Communication View. First, all of these theorists understand the reactive attitudes, or, more generally, blame, to be a form of moral address, communicating a message (most often construed as a demand) to their object. And second, these theorists take the standards for the propriety of the reactive attitudes, or, again, blame, to issue from the constraints on the intelligibility of communicating this message. Moreover, these two claims are logically related: the second claim, it seems, is supposed to follow from the first. The fact that the reactive attitudes are forms of moral address, in other words, is what makes it the case that the propriety of the reactive attitudes is a matter of the conditions for the intelligibility of the particular form of moral address they represent.

My worry about the Communication View ultimately concerns the relation between its two claims. Particularly, I suspect that either the Communication View's first claim is too strong, or it isn't clear how the second claim follows from it.

Consider the thought that blame isn't only a form of moral address. Blame, that is, might involve elements other than the communicative element on which the Communication View focuses. Many theorists, for instance, have argued that blame is a way of protesting the threatening claims expressed by others' actions (see Hieronymi 2001; Talbert 2012; Smith 2013). Others have taken blame to be a way of promoting socially desirable behavior (Smart 1961). If these theorists are right, then there might be more at stake in blaming someone than communicating messages.[7] And if this is the case, then it isn't clear why blame's communicative element should determine the standards for its propriety rather than, say, it's protest element. Indeed, Watson wrote that the reactive attitudes lose their point "as forms of moral address" if they target someone who is incapable of comprehending the demand they communicate. But does this mean that they lose their point entirely?

Of course, I don't mean to argue that blame in fact has these other elements. Nor do I mean to argue that these other elements are relevant to blame's standards of

[7] There is a slight complication here. Protest, one might think, seems like a sort of communication. Couldn't the Protest View, then, be subsumed under the Communication View? Not quite. There is a key difference: on the Communication View, the thought is that blame calls for some response from its object. On the protest view, though, protest doesn't seek any response in this way; rather, it is defensive—a way of standing up for oneself, often in the face of certain incorrigibility (cf. Talbert 2012, 105-107).

propriety. I merely wish to draw out a commitment of the Communication View. It seems the Communication View is committed either to denying that blame has these other elements or to holding that blame's communicative element is specially privileged to determine the standards for blame's propriety.[8] In either case, I take the Communication

[8] There is admittedly another way one might develop the Communication View. One might suggest that communication is just one among many elements of blame, all of which determine blame's standards of propriety. This approach could take two forms: (1) a disjunctive form, according to which one is an appropriate object of blame so long as one of blame's elements has application, and (2) a conjunctive form, according to which one is an appropriate object of blame only if all of blame's elements have application. If Communication theorists have had either of these views in mind, they haven't been entirely clear about it. Indeed, both of these views would make communication far less central to blame's standards than Communication theorists have suggested it is. Moreover, it would mean Communication theorists haven't provided a full account of blame's propriety. This is particularly damning in the case of (1), according to which someone could be an appropriate object of blame even if communication didn't have application. Nevertheless, one might develop the Communication View along the lines of (2). On (2), blame's communicative element would provide some (but not all) of the necessary conditions for blame's propriety. Unfortunately, I don't have the time or space to pursue this line of thought fully here. Some brief remarks must suffice.

To begin, it is worth noting that (2) needs some fleshing out. First, it faces questions about what elements of blame determine its propriety and about why *those* elements are the relevant ones. For instance, blame seems to have a regulative element; however, many have objected to the thought that blame's propriety should be a matter of whether one's blame would conduce to the regulation of other people's behavior. But second, and more importantly, one might question the conjunctive strategy itself. Why should blame be inappropriate if, say, its protest element has application but its communicative element doesn't? This relates to the point I make in the text above: if some of blame's elements have application, why should blame lose its point *entirely*? Macnamara 2015b suggests that blame's communicative element is a function of blame, and this provides a useful analogy. A screw gun might have the function of putting screws into objects and taking screws out of objects, but it doesn't seem inapt to use it for one of these functions and not the other. We might apply a similar thought to blame. It simply isn't clear why all of blame's elements must have application for it to be apt.

There is far more to say here. But what is the upshot for this paper? It seems possible to develop the Communication View in accord with (2). It is unclear, though, whether this is what Communication theorists have had in mind in their articulation of the Communication View. And there is some reason to think certain theorists haven't had (2) in mind. David Shoemaker (2007), for instance, is interested in both the necessary *and* sufficient conditions for blame's propriety. If one accepts (2), though, one cannot derive the sufficient conditions for blame's propriety from blame's communicative element alone. Watson (1987) also doesn't seem to have (2) in mind. He suggests the Communication View as a way of protecting Strawson's view from the incompatibilist about moral responsibility and determinism. Watson worries that Strawson's view of exemption has a gap in it and that this gap might be filled in by an incompatibilist condition on morally responsible agency. If this is Watson's motivation for proffering the Communication View, it suggests that he meant to entirely fill the lacuna in Strawson's account. If Watson had the conjunctive view in mind, however, he would only have partially filled that lacuna.

Still, if Communication theorists adopt a view like (2), then the Communication View might avoid the worry I raise for it in this paper. That worry, then, might best be cast as a worry for a particular construal of the Communication View, which certain theorists seem to adopt, and which

View to hold that blame is fundamentally about communication; thus, communication is normatively central to our blaming practices. Without such a commitment, it is unclear how the Communication View is licensed in drawing the standards of blame's propriety from blame's communicative element.

When the Communication View's first claim describes blame as a form of moral address, then, it is doing something more than describing just another feature of blame. It is positing communication as what is fundamentally at stake when we blame people: when the relevant communication is unintelligible, blame likewise no longer makes sense. In the next two sections, I hope to challenge this idea by suggesting that there is more at stake in our blame than communication. It is worth noting, though, that I don't take myself to definitively show that blame's communicative element *isn't* what determines blame's standards of propriety. I merely hope to show that is far from clear that the Communication View establishes that it *is*.

**3. Blame and Forgiveness**

I don't doubt, then, that blame has some communicative element. I do, however, have doubts that this element is as central to blame as the Communication View suggests. To see why, I now turn to the conceptual connection between blame and forgiveness. In this section, I suggest that any theory of blame's nature must elucidate our reasons for forgiveness. If blame is fundamentally about communication, then, this communicative nature should tell us something about our reasons for forgiving people. In the next section, I'll attempt to show that communication doesn't seem relevant to our reasons for forgiveness.

To start, we should get clearer on what forgiveness is. For my purposes, I mean to adopt as minimal a conception of forgiveness as I can. I thus suggest the following model: forgiveness involves changing one's perspective concerning blaming another person.[9] This model is meant to be flexible. The relevant change in perspective might be a matter of forswearing one's blame, judging that one's blame is no longer fitting or

others don't obviously repudiate. My worry, then, might also be taken to suggest that the Communication View should be defended along the lines of (2). As I've suggested, though, this raises another set of questions, including a question about the justification for the conjunctive view itself.

[9] More precisely, it involves changing one's perspective *away* from blaming the other person (as opposed, say, to coming to feel as though one *should* blame that person).

good, or ceasing to blame the other person entirely. I also intend to remain neutral on the meaning of "blaming another person." Of course, given the larger purposes of the paper it is perhaps most natural to follow Strawson and the Communication View in understanding blame to involve resentment or indignation, but there seems to be no reason to limit blame to this here. In any case, the point is simply that forgiveness minimally involves some change in perspective concerning blaming another person. It doesn't seem I'm in a position to forgive someone whom I don't already blame, or, at least, whom I don't already see as blameworthy. If my friend lies to me, for example, then it seems I can only forgive her if at least the thought of blaming her occurs to me; otherwise, there is, in a sense, nothing for me to forgive.[10]

Minimally, then, I take forgiveness to involve changing one's perspective concerning blaming another person. But this change in perspective cannot come about in just any way. Pamela Hieronymi, for instance, has pointed out that one hasn't genuinely forgiven someone if one simply takes a pill that brings about the change in perspective forgiveness involves (Hieronymi 2001, 530). Similarly, if the change in perspective comes about due to head trauma or amnesia, it doesn't seem as though there is forgiveness. In these scenarios, the change in perspective simply happens to the supposed forgiver. When we forgive another person, however, it doesn't seem we are passive in this way. Rather, forgiveness is a rational activity. As Jeffrie Murphy puts it, forgiveness is "the sort of thing one does for a reason." And this seems to be what differentiates it from merely forgetting, "which may just happen" (Murphy and Hampton, 1988, 15).[11]

If forgiveness is the sort of thing one does for a reason, though, this raises a question: What are the right kinds of reasons to forgive someone? Insofar as the change in perspective that forgiveness involves concerns blame, it seems like the relevant kinds of reasons will likewise concern blame. This is a start. But what considerations concerning blame will be relevant to the question of whether to forgive someone?

[10] Granted, assuming my friend is in fact blameworthy, there is, in another sense, something for me to forgive: her wrongdoing. However, my point here is simply that before I'm in a position to forgive my friend I have to start blaming her, or at least thinking about blaming her, in the first place.

[11] One might worry about the idea that forgiveness is something we *do*. Forgiveness—changing one's perspective—doesn't seem the same as, say, picking up a pencil. Here, I (and I take it Murphy and Hieronymi) have in mind something different from voluntary action. We *forgive* in the same sense that we paradigmatically *form* beliefs or emotions. Such doings are a result of our rational activity. They thus seem different from things that simply happen to us.

Prudential considerations are one candidate. It isn't uncommon for people to treat blame as something to avoid. This isn't surprising. Blame can have many negative effects.[12] It can be consuming—distracting the blamer from other important aspects of her life; it often means the fracture of a relationship, and an attending loss of support and camaraderie; and the stress and anxiety blame involves are unpleasant and might even negatively impact one's health. Are these sorts of prudential considerations reasons to forgive someone?

It may seem so. We could imagine a therapist, for example, urging his client to forgive her father because of the toll her blame takes on her. I think, though, that treating these considerations as reasons to forgive is misleading. They are certainly reasons to think one's blame is bad for one, and they might be reasons to become a forgiving person (Roberts 1995); but these prudential considerations don't seem like reasons for forgiveness itself. We could imagine, for instance, the above client offering the following rejoinder to her therapist: "I know that my blame is hard on me. And I wish I could forget it. But I can't possibly forgive my father. He was cold and manipulative, and he never showed any remorse for his behavior." Here, the client seems to be suggesting that prudential considerations aren't enough; the right sorts of considerations for forgiveness are missing. This seems like a fitting response to her therapist's suggestion. Moreover, if I wrong my friend and she blames me, and if I desperately want her forgiveness, trying my best to make amends, it doesn't seem I'd be satisfied if I found out she "forgave" me simply due to the deleterious impact her blame had on her. In fact, if I found out my friend's reasons were purely prudential in this way, I wouldn't feel forgiven. Again, the point isn't that prudential considerations cannot factor into forgiveness at all—they might be reasons to become ready to forgive people—but it seems they aren't reasons for forgiveness itself. Or to put the point another way: prudential considerations might be reasons to bring about the change in perspective forgiveness involves, they might be reasons to take Hieronymi's "forgiveness" pill, for example; however, they don't seem to be reasons for that change in perspective itself.

[12] Of course, this isn't to suggest that blame isn't nevertheless valuable. In fact, I worry the negative effects of blame are most often overplayed at the expense of its positive value.

To make progress on our question, then, we should consider the sorts of reasons for the change in perspective that forgiveness itself involves. This change in perspective concerns blame. It is natural to think, then, that the kinds of reasons relevant to forgiveness will have to do with why we're blaming (or preoccupied by the idea of blaming) the person in the first place. But we have to be careful here, for there is a set of reasons that have to do with why we blame people that cannot factor into genuine forgiveness. As Hieronymi explains, there are three interrelated judgments that undergird our blame but that cannot be given up in forgiving another person: (1) the judgment that the person's action was wrong; (2) the judgment that the person is the kind of agent who is morally responsible for her actions; and (3) the judgment that you, the person wronged, shouldn't be treated that way (Hieronymi 2001, 530). All of these judgments concern the culpability of the person being blamed, according to Hieronymi. And to give any of them up is to excuse the person's behavior, not to forgive it. Considerations that bear on any of these three judgments are not reasons to forgive; they are reasons not to blame.

Some aspects of why we blame thus don't help us understand the reasons for forgiveness. Nevertheless, there is more to say about why we blame people than simply that the person is culpable. Here, I suggest we take a lesson from Hieronymi's account of forgiveness, which she develops to solve a challenge similar to the one we're facing.[13] Hieronymi's account of forgiveness begins with an account of blame's nature. For Hieronymi, blame—resentment—"protests a past action that persists as a present threat" (Hieronymi 2001, 546). More specifically, blame challenges threatening claims expressed by wrongful actions. And, according to Hieronymi, this aspect of blame's nature suggests a judgment underlying our blame: "the event in question makes a threatening claim" (548). Hieronymi argues that this is the judgment that gets revised when we forgive someone. When we forgive someone, we revise our judgment about whether the relevant wrongdoing still makes a threatening claim. So, for Hieronymi, the reasons for

[13] Hieronymi puts the challenge like this: any account of forgiveness must be both articulate and uncompromising. It must be articulate because it must articulate the "revision in judgment" that forgiveness involves. And it must be uncompromising because the revision in judgment it articulates cannot involve changing any of the three judgments concerning culpability. My thought in this section is very much indebted to Hieronymi's paper.

forgiveness are those considerations that bear on the question of whether some action's claim is still threatening.

While I find Hieronymi's account of forgiveness appealing, I don't mean to advocate it here. Rather, I hope to draw a general lesson from it. Hieronymi's solution suggests that the change in perspective that forgiveness involves will concern blame's nature—what blame is fundamentally *about.* And this makes sense. It seems like our reasons for forgiveness, our reasons for changing our perspective concerning blaming another person, should have to do with what's at stake in our blame in the first place. In this way, the nature of forgiveness appears intimately tied to the nature of blame, and any account of forgiveness will involve a corresponding account of blame. Moreover, because forgiveness concerns what's at stake in our blame, it seems intuitive that our reasons for forgiveness might be tied to the aspect of blame relevant to exemption; our reasons for exemption, after all, also seem to be a matter of what's at stake in our blame in the first place.

We can therefore judge accounts of forgiveness on the basis of the account of blame they rely on. We might, for instance, argue that Hieronymi is wrong about blame being protest. And this would suggest we must account for the nature of forgiveness differently than Hieronymi. But the conceptual connection between blame and forgiveness is not a one-way street. If an account of blame's nature cannot explain the reasons for the change in perspective that forgiveness involves, it suggests that the account is at least incomplete, if not wrongheaded. This puts a constraint on our theorizing about blame: any account of blame's nature must elucidate our reasons for forgiveness.

**4. Communication and the Reasons for Forgiveness**

I've suggested, then, that any account of blame's nature must elucidate our reasons for forgiveness. If I'm right about this, then I believe we have reason to worry about the Communication View. The Communication View, remember, puts blame's communicative element at the center of our blaming practices. It doesn't seem, though, that blame's communicative element elucidates our reasons for forgiveness. This suggests there is more at stake in blame than communication, and perhaps even that communication isn't fundamentally what's at stake in our blame.

Consider a paradigmatic case of forgiveness. Your friend lies to you about something that, in the grand scheme of things, isn't very important. When you find out about her lie, you resent your friend, and you distance yourself from her; maybe you even tell her off. Finally, your friend, feeling guilty about her transgression, apologizes. You forgive her. To fit with the Communication View, which understands blame in terms of the reactive attitudes (particularly resentment), let's say that your forgiveness here is the forswearing of your resentment.[14] How might blame's communicative nature elucidate your reasons for forgiving your friend?

Presumably, the reason for forgiving your friend in this situation has to do with her guilt and her apology. As a first proposal, then, your reason for forgiveness could be that your friend's guilt and apology are evidence that she "heard," so to speak, the message your resentment communicated. But this doesn't seem right; it wouldn't even require that your friend apologize or feel any guilt. We can hear messages, after all, without responding to, or even caring about, them.

This points the way towards a second proposal. Perhaps your reason for forgiving your friend is that she felt guilt and apologized *because* of the message (e.g., the demand or call) your resentment communicated. It wasn't just that your friend heard the message, then, but also that she responded to it. This seems better. However, it still doesn't seem quite right. That your friend merely responded to your resentment with guilt and apology seems like an odd reason to forgive her. This is easier to demonstrate with apology: if your friend responds to your blame with an insincere apology, it doesn't seem like this is a reason to forgive her. We can make the same point, albeit more fancifully, with guilt. Say you blame your friend but she isn't moved by your blame at all; she simply doesn't think she owes you anything (despite that she wronged you). She knows, though, that you won't be satisfied unless she feels guilt, and she wants to satisfy you for self-interested reasons. She undergoes hypnosis to feel guilt. Here, it doesn't seem like the fact that your friend responded to your blame with guilt is reason to forgive her.

[14] The idea that forgiveness is forswearing one's resentment is perhaps the most common conception of forgiveness in the literature. It is worth noting that it fits my minimal conception of forgiveness from the previous section: forswearing one's resentment can be understood as a change in perspective (forswearing) concerning one's blame (resentment).

It thus appears that we're interested in more than mere response when we blame people. It isn't the guilt and apology themselves that matter but, rather, what they signify. This suggests a third proposal on behalf of the Communication View. Perhaps you forgive your friend because she is *receptive* to your blame's message. In other words, your friend responds to your blame with guilt and apology because she feels the force of your blame and the message it communicates. Your blame draws her attention to the fact that she harmed you in a particular way, and this brings her to genuinely apologize and feel guilty. This appears to be the most promising way for the Communication View to elucidate our reasons for forgiveness.

But we might ask: What is the significance of this receptivity? It seems like this receptivity is significant because it represents some sort of goodwill or regard for your person and, perhaps also, for morality.[15] Indeed, this goodwill and regard seem to be the significance of someone apologizing for or feeling guilty about wronging you on some particular occasion: they didn't mean it; they don't think that is how you ought to be treated. But this suggests that blame's communicative element isn't playing a role in your reason for forgiveness. To bring this out, consider the following two candidate reasons for that forgiveness:

(A) Your friend displayed genuine goodwill and regard for you because your blame called for it

(B) Your friend displayed genuine goodwill and regard for you because she has genuine goodwill and regard for you.

If your blame's communication is truly playing a role in your reasons for forgiving your friend in our paradigm case, it seems your reason should be of (A)'s form. However, (B) seems closer to the kind of consideration that normally justifies our forgiveness. Your friend's receptivity to the message your blame communicates is simply indication of (B). Consider, for instance, a modified version of our paradigm case. Say you find out about your friend's lie because, racked with guilt, she comes clean to you about it, apologizing profusely. In this case, your friend's guilt and apology aren't responses to your blame at

[15] For an account of forgiveness in a similar (Strawsonian) vein, see Martin 2010, 541-546.

all. Nevertheless, you still face a decision about forgiving your friend. And while your friend shows you goodwill and regard, it seems doubtful that, in this scenario, your reasons for forgiving your friend would have anything to do with blame's communicative nature. What seem important are your friend's regard for you and her goodwill itself. Our original case doesn't seem different. Your blame may trigger your friend's regard and goodwill, but it seems to be the regard and goodwill itself that matters, not the fact that the regard and goodwill come about because of your blame. It thus doesn't seem that blame's communicative element elucidates our reasons for forgiveness.

**5. Conclusion**

If I'm right that blame's communicative element doesn't elucidate our reasons for forgiveness, what does this mean for the Communication View? To establish its claim that the standards of propriety derive from blame's communicative element, the Communication View seems to take blame to fundamentally be about communicating messages. If blame's communicative element doesn't elucidate our reasons for forgiveness, though, this suggests that there is more at stake in our blame than communication. And if there is more at stake in blame than communication, it is unclear why communication is specially privileged to determine the standards for blame's propriety. Above, for instance, we saw that our reasons for forgiveness might have to do with people showing us goodwill and regard. If a concern for people showing us goodwill or regard is at stake in our blame, the capacity to show goodwill or regard could simply be what determines whether an agent is exempt from our blaming practices. Of course, I don't mean to put forth an account of blame's standards here. Nor do I mean to suggest that blame's communicative element is irrelevant to those standards. My point is merely that it isn't obvious communication is as important as the Communication View makes it out to be—either to what is at stake in blame or to blame's standards of propriety.

More work needs to be done here, particularly in relating blame's nature to blame's standards of propriety. How, precisely, are these two things related? Why is it that the standards of propriety should derive from one aspect of blame rather than another? I'm not in a position to pursue these questions here. However, I hope to have shown one possible way forward. Thinking more about how blame fits into a larger picture of moral life and moral relationships, including the way it is conceptually

connected to other phenomena that our moral lives and relationships feature, might help us get a better sense of blame's nature, especially what's at stake when we blame someone.

Works Cited

Austin, J.L. 1975. *How to Do Things with Words*. Cambridge, MA: Harvard University Press.

Darwall, S. 2006. *The Second-Person Standpoint: Morality, Respect, and Accountability.* Cambridge, MA: Harvard University Press.

Fricker, M. 2016. What's the point of blame? A paradigm based explanation. *Nous* 50(1): 165-183.

Hieronymi, P. 2001. Articulating an uncompromising forgiveness. *Philosophy and Phenomenological Research* 62(3): 529-555.

Macnamara, C. 2015a. Blame, communication, and morally responsible agency. In *The Nature of Moral Responsibility: New Essays.* Edited by R. Clarke, M. McKenna, and A. Smith. Oxford: Oxford University Press: 211-236.

Macnamara, C. 2015b. Reactive attitudes as communicative entities. *Philosophy and Phenomenological Research* 90(3): 546-569.

Martin, A. 2010. Owning up and lowering down: the power of apology. *The Journal of Philosophy* 107(10): 534-553.

McKenna, M. 2012. *Conversation and Responsibility.* Oxford: Oxford University Press.

Murphy, J. and Hampton, J. 1988. *Forgiveness and Mercy.* Cambridge, UK: Cambridge University Press.

Roberts, R.C. 1995. Forgivingness. *American Philosophical Quarterly* 32(4): 289-306.

Scanlon, T.M. 2008. *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, MA: Harvard University Press.

Shoemaker, D. 2007. Moral address, moral responsibility, and the boundaries of the moral community. *Ethics* 118(1): 70-108.

Smart, J.J.C. 1961. Free will, praise, and blame. *Mind* 70: 291-306.

Smith, A. 2013. Moral blame and moral protest. In *Blame: Its Nature and Norms*. Edited by D.J. Coates and N. Tognazzini. Oxford: Oxford University Press: 27-48.

Strawson, P.F. 1962. Freedom and resentment. *Proceedings of the British Academy* 48: 1-25. Reprinted in *Perspectives on Moral Responsibility*. Edited by J.M. Fischer and M. Ravizza. Ithaca, NY: Cornell University Press: 45-66.

Talbert, M. 2012. Moral competence, moral blame, and protest. *Journal of Ethics* 16(1): 89-109.

Watson, G. 1987. Responsibility and the limits of evil: variations on a Strawsonian theme. In *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*. Edited by F. Schoeman. Cambridge, UK: Cambridge University Press: 256-286.

Watson, G. 2011. The trouble with psychopaths. In *Reasons and Recognition: Essays on the Philosophy of T.M. Scanlon.* Edited by R.J. Wallace, R. Kumar, and S. Freeman. Oxford: Oxford University Press: 307-331.

David Shoemaker
Tulane University

**Bio:** David Shoemaker is a Professor in the Department of Philosophy & Murphy Institute at Tulane University. He is the author of *Responsibility from the Margins* (OUP 2015), *Personal Identity and Ethics: A Brief Introduction* (Broadview 2009), and numerous articles on agency, responsibility (criminal and moral), personal identity, ethics, and moral psychology. He is the general editor of the OUP series *Oxford Studies in Agency and Responsibility* as well as the organizer of the associated biennial New Orleans Workshop in Agency and Responsibility (NOWAR).

# THE FORGIVEN

According to most responsibility theorists, for my response to your offense against me to count as *blame*, it must include an emotional component, and this is almost always thought to be resentment.[1] To understand the nature, reach, and resolution of blame, then, it has seemed natural to start theorizing by analyzing resentment. The most familiar story about it goes as follows. Resentment is an angry emotional response with a constitutive cognitive component, typically the judgment that one was wronged.[2] If (and only if) the offender did wrong one, one's resentment of him is appropriate. Now to the extent that one cares about how others treat one, and one cares about morality generally, when someone violates morality's tenets with respect to one, one's triggered resentment expresses that caring (see, e.g., Franklin 2013; Wallace 2010: 323-24; Wallace 2013: 230). This expression is blame.

Unfortunately, there are multiple analyses of resentment, and they yield multiple disagreements. Building from an analysis of resentment, some take blame's function to be *protest* (Hieronymi 2001 and 2004; Smith 2013; Talbert 2012), whereas others take it to be a

[1] It's easier to give the main exceptions to what is a long list: Sher 2006 and Scanlon 2008.

[2] See, e.g., Wallace 1994; D'Arms and Jacobson 2003; Darwall 2006; McKenna 2012; and many others.

kind of *demand* (Darwall 2006; Strawson 2003; Wallace 1994; Walker 2006; Watson 2004). And even within these interpretations, there is little agreement over, for example, what precisely blame protests against, or what it is a demand for. And the disagreements generated by the front-end method of investigating blame multiply even more when we are told what it implies for blame's resolution at the back end in forgiveness. If we cannot clearly determine what blame's function is, or what its demand is for, say, then we are going to have serious problems determining what resolves it.

What has gotten us into this mess is that blame theorists have put all of their eggs into the resentment basket on the front end of the blaming exchange. They have assumed that resentment is the paradigm blaming attitude, and then they lean on their analysis of it to work out the nature of the whole rest of the exchange. But as these problems suggest, we cannot squeeze so much out of resentment; indeed, it has been wrung dry. In this essay, therefore, I take a fresh approach to these issues by starting in on the *back end* of the blaming exchange, namely, the point at which blame is appropriately withdrawn via forgiveness.[3] In particular, I will focus on what it takes to be successfully *forgiven*. This approach reveals fruit we cannot see by starting on the front end of the exchange with resentment, including: (a) why blame functions as a demand, not a protest, (b) what blame's demand is for, (c) why judgment is superfluous to the blaming exchange, (d) why resentment has been the wrong blaming attitude to lean on all along, and (e) what the paradigm blaming attitude is instead. This is a lot of fruit, so let's get picking.

[3] In Victoria McGeer's psychologically insightful "Civilizing Blame," we see a glimpse of this sort of argumentative strategy, more generally applied to "what makes our blaming emotions go away" (McGeer 2013: 174). In some important respects, her focus is different than mine, as she talks about blaming attitudes that resolve in light of excuses and exemptions, not forgiveness. But she does at least mention forgiveness and remorse as among possible blame-resolvers, despite not pursuing their character in any detail, and she does focus on anger rather than resentment in blame, which we will eventually see that I'm all in favor of (it's just that I see an argument for doing so coming out of the back-end of our blaming exchanges, via a focus on forgiveness, whereas McGeer discusses anger at the front end of blaming exchanges insofar as it is the evolutionary ancestor of contemporary blame).

*The Forgiveness Puzzle*

There is a standard puzzle about forgiveness: How is it that the forgiver could reject her blame as no longer appropriate without *excusing* the offender (Hieronymi 2001: 530)? Forgiving must be compatible with simultaneously maintaining the view that the offender still offended against one, but how could this result be effected? Perhaps the most influential resolution of this puzzle comes from Pamela Hieronymi. On her view, when you commit a moral offense against me, I actually make multiple judgments, but the two that matter are these: (a) a judgment that you (responsibly) wronged me; and (b) a judgment that, in wronging me, you have expressed the false claim that I can be treated in this way, that such treatment is acceptable (Hieronymi 2001: 546).[4] Such a false claim, as long as it remains unaddressed, constitutes an ongoing *threat* to me. And how do we come to see this point? By following the familiar methodology, or, as she puts it, "[W]e need to delve more deeply into the attitude of resentment" (Hieronymi 2001: 545). She interprets it as the emotional *protest* constituted by my (b)-judgment, so it is appropriate just in case that (b)-judgment (that you've made the threatening false claim) is itself true. Consequently, when you genuinely apologize and renounce your deed, and I forgive you, my resentment is no longer appropriate insofar as your renunciation has wiped out the threatening false claim to which my resentment is a response, and so it has rendered my resentment's constitutive (b)-judgment false. Nevertheless, my *(a)*-judgment—that you responsibly wronged me—is still true. But insofar as the appropriateness of resentment is not a matter of the truth of the (a)-judgment—I can judge that you responsibly wronged me without (appropriately) resenting you—it is possible for me both to maintain that (a)-judgment and coherently forgive you (by withdrawing the resenting (b)-judgment as no longer justified).

[4] Other judgments include that the offender is a fellow member of the moral responsibility community and that he's worth being upset by (Hieronymi 2001: 530).

This is an elegant solution, made more plausible to many given that it also captures what they think is a crucial component of resentment—the paradigm blaming emotion—namely, it is constituted at its core by a *judgment* (even though there is disagreement about what that judgment is). Call this view *judgmentalism*. I believe judgmentalism is false. My intention is to paint a very different—and much more determinate—picture of both blame's nature and its aim that is revealed when we start at the other end of the blaming exchange.

*Resolving Blame via Forgiveness*

What makes an instance of forgiveness normatively successful? In other words, what about an offender's apology, say, effects an appropriate transition in the forgiver from holding it against the offender to not holding it against her, while nevertheless not lapsing into excusing her?

Psychological studies reveal that the degree to which successful forgiveness is most likely effected depends on the degree to which several distinct features of apologies are in place, among them admitting fault, admitting damage done, and offering to make amends (see Schmitt et al. 2004; Zechmeister et al. 2004; and Dill and Darwall 2014: 51). But by far the most significant predictor of forgiveness is expressed (or perceived as sincere) remorse (Davis and Gold 2011: 392). So what is remorse? There are—no surprise—competing moral psychological accounts, many putting remorse in the same camp as either guilt or shame (or a hybrid of both). This fact might threaten to make any conclusions of my project as indeterminate as the resentment-first project. But I think there is a characterization of remorse that is clearly superior to others, and it is based on the idea that remorse is a distinct emotional syndrome with a distinctive motivational impetus.

To explain, emotions have a syndrome, typically consisting in felt affect, associated thoughts, and an action tendency. For identification and differentiation purposes, the most important element of this syndrome is the emotion's action tendency (Scarantino 2014: 168-183), which is a state "of readiness to execute a given kind of action," one "defined by…[the]…end result aimed at" (Frijda 1986: 70; quoted in Scarantino 2014: 169). Given this influential account of emotions, I think the characterization of remorse articulated by Alan Thomas is most plausible: "Remorse, by contrast with either shame or guilt, [is a response to] the destruction of value rather than the infringement of standards of right and wrong" (Thomas 1999: 130). In cases of interpersonal offenses, guilt motivates one to repair the relationship, or right the wrong. When one is ashamed for what one did to another, one is moved to hide, either from the gaze of the wronged party or enforcers of the relevant standards. But remorse itself may be without either action tendency. What distinguishes it from both guilt and shame is that it tends one toward reflection on—or often wallowing in—some disvaluable state of affairs that one caused. It moves one to relive the relevant events over and over, bemoaning the loss one caused.[5] But there is also, in its most powerful and pure examples, a sense of impotence to remorse, the sense that *all* one can do now is reflect on or wallow in the damage. And, furthermore, where the value damaged is a fellow moral creature, one's contemplation of this lost value reflects one's identification with that fellow (cf., Deigh 1996: 50). Of course cruel people may be moved to identify with their victims and delightfully reflect on what they did to them in an utterly non-remorseful way. So we must make sure to incorporate into our account the *painfulness* involved in this particular ruminative activity. My remorse for something I did to

[5] And so notice that remorse doesn't necessarily involve thoughts of having done otherwise. It could be that one appropriately feels remorse for having done the only thing one could have done, where it nevertheless caused a significant loss of value. Thanks to Dan Tigard for discussion.

you, then, in its most resonant form, consists in my painful emotional response to my recognition of having caused an irremediable loss of value in you, a response which constitutively involves my being moved to reflect on (perhaps again and again) what I did.

I take this to be an uncontroversial characterization of remorse. If it does not capture all such instances of the folk emotion term,[6] then I am happy to stipulate it simply as an unnamed, but very familiar, emotional syndrome (although I will continue to use the term "remorse" to describe it).

Nevertheless, an immediate question arises: If I felt no pain on actually damaging your value at the original time of action (which is usually the case), then why should I feel pain on later imaginatively revisiting what I did? If my experience of it didn't hurt the first time, why should my memory of that experience bring the pain? The answer is that I am not simply replaying that action in my head as it was experienced by me: the imaginative revisitation is not a mere memory. Rather, I must now be seeing what I did from a different perspective, namely *yours*.[7] And not only that (to avoid the cruelty counterexample), I must also be open to feeling some approximation of how you felt upon being the victim of my bad treatment. The painful feeling of remorse is to a great extent a simulacrum of how you feel about the loss I caused, as

[6] Some may claim, for instance, that I can sensibly feel remorse over many deaths caused by a Nepalese earthquake, or for a lost language. I am not an emotion-term chauvinist, so I am fine with granting the use of the term "remorse" to cover such feelings. As I note in the text, all I want to do is carve out the more limited boundaries of a very familiar emotional syndrome, regardless of its label.

[7] Why couldn't I be experiencing it from the perspective of someone else, a neutral bystander? It's not clear how the constitutive painfulness of remorse could be generated in that case. If you are the bystander, after all, you may be happy at the value lost—perhaps you had a bet on it. So what must be the case is that remorse involves reflection on the value lost from the perspective of the person whose loss it is. Thanks to Nick Sars for discussion.

from your perspective. In other words, what enables the relevant sort of remorse in interpersonal cases is what I will call *empathic acknowledgment*.[8]

*Acknowledgement and Judgment*

Stipulate, then, full empathic acknowledgment on my (the offender's) part. At that point, resentment tends to feel disarmed and its suspension is appropriate. The conditions for paradigm forgiveness, in other words, have been met.[9] But why? Here we may seem to face a Hieronymi-induced problem, for there is nothing about full empathic acknowledgment, in and of itself, that would render false the offender's *judgment* that the victim can be treated poorly, and so nothing about the offender's acknowledgment that could appropriately disarm the victim's resentment of him. After all, empathic acknowledgment simply consists here in an emotional *perceptual stance*, a simpatico appreciation from the inside of just how it (must have) felt for the victim to be treated in the way she was.[10] But one's taking up an emotional perceptual stance is irrelevant to the status of one's threatening (false) claims about other people. And remorse is simply a painful emotion that tends to motivate its experiencer to reflect on and ruminate over the loss he or she caused. But this too has no engagement with one's threatening claims about how others may be treated. Instead, it looks as if blaming protest of those claims could be resolved—and

[8] See also Thomas's remarks: "[T]he problematic demand for reparation in the case of remorse seems to reflect our demand to the agent that he or she do more than recognise that he or she brought about the bad state of affairs through his or her agency. Rather, that the agent should acknowledge that he or she understands what he or she has done—some writers have spoken in this connection of a deepening sense of the 'moral meaning' of an agent's action" (Thomas 1999: 133).

[9] This may be a more controversial point than I think it is. Unfortunately, while I do defend it a bit below, I will not give much positive support for it, other than to say that it strikes me as phenomenologically accurate. Call it prima facie plausible, then. In addition, this formulation allows for plenty of non-paradigm cases, such as elective forgiveness. I'm just articulating one central case in which everyone would agree the appropriateness conditions for forgiveness have been met.

[10] I explain this idea in much greater detail in Shoemaker 2014 and 2015: 99-100.

thus successful Hieronymian forgiveness could be achieved—only via the offender's explicit *repudiation* of the threatening claim.

Nevertheless, repudiation is no constitutive part of, and on its own isn't enough to establish the conditions of, many paradigm cases of normatively successful forgiveness. Suppose I am texting while driving and run over your dog because of my lack of attention to the road. Once I see the damage I have caused, I am overwhelmed with remorse. I too have a dog, and I can imagine just what it must have been like for you, and I feel your pain at the loss of value I caused from your perspective. It looks appropriate for you to abandon resentment in favor of forgiveness just as soon as you have witnessed my own sincere emotional devastation in light of what I did.[11] It is obvious that I clearly and truly "get" what I did, and that may well be sufficient for your appropriate forgiveness, even without my repudiation of it. Indeed, there are lots of cases like this. When one's emotional devastation in light of one's realization of one's offense is expressed, it serves as a robust epistemic marker for the blaming agent, and so appropriately salves the blamer's negative emotions.

Repudiation alone may also be insufficient for normatively successful forgiveness. Suppose I have broken a promise to help you move this morning. When you angrily make me aware of the fact that I was supposed to help you, my immediate response is to tell you, in a flat voice, "Yeah, I didn't really feel like coming over this morning but I agree that that violates moral standards. Nevertheless, I resolve not to violate those standards in the future. I know full well that you aren't supposed to be treated in this way." Suppose these are sincere assertions, and you recognize them as such. Still, you may well have the sneaking suspicion that

[11] In 2011, Patricia Machin's husband was hit and killed by a careless driver. In a note forgiving the utterly distraught man, she wrote, "However bad it was for me, I realize it was 1,000 times worse for you." From *The Telegraph*, February 20, 2013, URL: http://www.telegraph.co.uk/news/politics/david-cameron/9883398/Humbled-by-the-courage-of-those-who-forgive.html.

forgiveness has not been earned, and you would be right. My merely withdrawing a claim about how you can be treated and resolving to do better in the future, however sincere, allows for the possibility that I don't yet really *get it*, that I don't *feel* it, that I don't yet appreciate what it is I put you through. I could also repudiate my "false claim" with a variety of insulting attitudes, including condescension or a grudging irritation. What would be missing in each case is a transition to repudiation via full empathic acknowledgment.[12]

Cases like these also reveal that someone might do something for which forgiveness could be appropriate without her actions even expressing a false claim that "you can be treated in this way, and that such treatment is acceptable" (Hieronymi 2001: 546). My texting-while-dog-killing does not make such a claim, for instance, as I hadn't even acknowledged you (or your dog) were there to be threatened. Indeed, emotionally wrought acknowledgment often gets us to realize that, while the offender may have acted *as if* she had made such a claim, she really hadn't; instead, she was just blithely unmindful.[13] But you may forgive her offending obliviousness while still maintaining that she indeed performed an oblivious offense (so you do not merely excuse her).

I take it, then, that appropriate forgiveness is a function, most fundamentally, of sincere empathic acknowledgment, not repudiation. But, as suggested above, acknowledgment—a perceptual empathic stance—does not answer to or engage with judgment. And this is true of any relevant judgment, not just Hieronymi's identified resenting-judgment. The function of blame that forgiveness fittingly discharges is thus not protest (at least of the sort previously

[12] Note that the requisite phrase for confessing Catholics is "Forgive me, father, for I have sinned," not "Forgive me, father, for I have repudiated my sinning." What the confessor has to do is fully describe the sin at that point, so as to *acknowledge* what she did in order to effect forgiveness.

[13] More controversially, I believe that we forgive for a much wider range of offenses than wrongings or wrongdoing, including failures to live up to certain expectations and emotional insensitivity. I lack space here to make this case, however.

defended). Rather, being a function of a stance enabled by the psychological action of perspective-taking, acknowledgment seems most directly a satisfactory response to a *demand*.[14]

*The Blaming Attitude and Its Demand*

So what is that demand for? When we start at the front end of the blaming exchange with resentment as our paradigm blaming attitude, we get numerous conflicting possibilities.[15] But starting at the back end with forgiveness reveals why resentment has been the wrong attitude to focus on all along. Nearly everyone these days agrees that resentment is a "cognitively sharpened" version of anger (D'Arms and Jacobson 2003: 143), sharpened by a judgment (typically that one was wronged). But the empathic acknowledgment grounding normatively successful paradigm forgiveness does not provide the grounds for revising or withdrawing a (blame-related) judgment.[16] So even when acknowledgment does resolve resentment via forgiveness, it could only resolve, not resentment's judgment component, but its *emotional* component, namely, its anger. (And as it resolves a kind of anger we have toward other agents, call it *agential anger*.[17])

Now recall that emotions have a triple syndrome—felt affect, associated thoughts, action tendency—and their identifying feature is their action tendency. What, then, is agential anger's

[14] Why couldn't it be a response instead to a desire, hope, or wish that the offender acknowledge what he did? It could, but I doubt it, primarily because the phenomenological character of blaming has the forceful feel of a demand and not any of these other attitudes. Interestingly, in cases of private, unexpressed blame, the phenomenological character may sometimes feel most like mere desire or hope. But in the transition from unexpressed to expressed (or active) blame, one's desire that the offender acknowledge what he did also seems clearly to transform into a *demand* that he do so. Indeed, what would be the point of communicating one's mere desire or hope in this way?

[15] And Macnamara 2013 has compellingly shown that each is actually quite problematic on its own.

[16] Except perhaps the judgment that the wrong remains unacknowledged, but this would be a very poor candidate for blame's constitutive component.

[17] Agential anger is distinct from a kind of *goal-frustration anger*, which can be produced by all sorts of (agential and non-agential) events, in virtue of its distinctive action tendency, which I am about to explore in the text.

action tendency? It is a type of anger that has long been thought (since Aristotle) to be a response to *slights* that contains the action tendency for *revenge*. This isn't quite right, however. While the motivational impulse to revenge is sometimes included in specific bouts of anger, this impulse is actually just a dramatic method for carrying out its more fundamental action tendency, namely, the impulse to *communicate the anger*. An argument for this view comes from consideration of pairwise cases, one in which one's angry revenge for a wrong is delivered without the wronging agent ever knowing that one was its angered source, versus one in which the same revenge is delivered with the successful communication that one was its angered source.[18] Only the latter feels like fully discharged anger.[19]

Let us, then, bring together several previous points: Blame's fundamental attitude, as revealed by paradigm cases of normatively successful forgiveness, is agential anger, whose action tendency is to communicate a demand for empathic acknowledgment of what the offender did, a demand appropriately resolved by said acknowledgment, which typically enables a simpatico (painful) experience of how he made the victim feel. As there are no judgments necessary to any part of this process and exchange, resentment (to the extent that a judgment is a constitutive component) is also not necessary to it.[20]

[18] See Shoemaker 2015: Ch. 3.

[19] And communication is more than expression, of course. I may yell at you all day long, but if you have your noise-cancelling headphones on, the mere expression of my anger won't do a thing to discharge it. Rather, what is required is your uptake of my attempted communication. This is another feature I take to be missing in the protest theory of blame.

[20] For greater defense of this claim, I steer the reader to Deigh 2011 and Shoemaker 2015: 88-89. Note that I am not denying that resentment is often a blaming response. Rather, I am saying both (a) there may be plenty of paradigmatic blaming responses that don't include resentment (as they are instances of mere agential anger), and (b) even when a blaming response does consist in resentment, it is not its constitutive judgmental component that makes it an example of blame; it is rather its (agential) anger component.

*The Aim of Demanding Acknowledgment*

Our final question, then, is this: Why does agential anger demand *acknowledgment*? After all, wouldn't it make more sense to demand compensation, say, for the harm caused? In response, consider Dustin Hoffman's character Ratso Rizzo, whose classic angry (and purportedly improvised) response in *Midnight Cowboy* to the taxi driver who almost hits him as he crosses a New York street goes as follows: "I'm walkin' here! I'm *walkin'* here!" Notice that there just is no harm here to compensate. So what is his anger's communicative point? It aims for the driver to register, to take seriously, Rizzo. *I'm* walking here, indeed. But what Rizzo demanded—acknowledgment—was exactly the same thing the driver should have done *before* their exchange, namely, register the fact of Rizzo's presence. The driver's offense consisted precisely in a failure of acknowledgment. But as such, it makes perfect sense that Rizzo would demand exactly what he didn't get before.

Anger demands of offending agents exactly what we implicitly demand of them pre-offense: due regard. This is just a demand that we be taken seriously, *that we be acknowledged*. Thus when you fail to take me sufficiently seriously in causing some offense, I will tend to demand in my response, via my anger, that you at least *now* do so. But while we demand acknowledgment both pre- and post-offense, the content of what we demand to *be* acknowledged differs. Pre-offense, I demand what I demand of everyone: that you take me and my ends sufficiently seriously. I count, and to the extent that I may be affected by what you do (or what your attitudes are), your expressed attitudes ought to reflect that fact. Post-offense, though, when it is clear that you have failed to take me sufficiently seriously, my anger demands that you acknowledge how you made me feel in *not* having properly acknowledged me pre-offense.

Paradigm cases of normatively successful forgiveness are those in which the terms of this demand have been sufficiently met.

What would such acknowledgment achieve? Acknowledgment may have a variety of positive future effects.[21] But what it achieves most fundamentally and immediately is the restoration of a kind of *normative equilibrium* between the blamer and the blamed that was upset by the blamed's initial and then persisting lack of (sufficient) acknowledgment. We all expect to count sufficiently in each other's deliberations as we make our way through the world. When you disregard me, I fail to count (sufficiently) for you. In empathically acknowledging me post-offense, you restore me to (what I take to be) my rightful normative place, as someone of significance in your practical deliberations and emotional life. Agential anger accomplishes its aim when such acknowledgment occurs, and to the extent that this is the ground for normatively successful forgiveness, empathic acknowledgment provides the conditions for anger's appropriate dissolution.

### *Resolving the Puzzle of Forgiveness*

Finally, we can now see how this non-judgmentalist, demand-based account of blame resolves the puzzle of forgiveness. When the goal of the agential anger's action tendency has been successfully met, one's agential anger is in fact no longer appropriate, insofar as the offender's empathic acknowledgment makes it the case that there no longer exists the slighting lack of acknowledgment to which it responded. My slighting you consists in my putting us into normative disequilibrium. We remain in a state of disequilibrium until it is corrected. The sort of empathic acknowledgment rendering forgiveness appropriate restores normative equilibrium. Where there is no normative disequilibrium, no agential anger is appropriate.

[21] See McGeer 2013 for a discussion of possible forward-looking features of blame.

Nevertheless, there *was* normative disequilibrium—a slight—and slights render agential anger appropriate. In cases of an offender's genuine empathic acknowledgment, then, the forgiver's anger is no longer fitting in virtue of the forgiver's having successfully gotten what she demanded from the offender, but the forgiver can still view what the offender *did* as a slight, and so view the offender as having *merited* agential anger, and so as having been *responsible* for the slight. The adherent of my account may thus distinguish excuse and forgiveness-inducing empathic acknowledgment as follows: the former makes agential anger inappropriate in virtue of its revealing that there was no slight; the latter makes agential anger inappropriate in virtue of its revealing that what was a persisting slight *is no more*. While a slighter can never make it the case that he did not slight the victim and so brought about normative disequilibrium with the slighted party, he can at least restore normative equilibrium between them by heeding agential anger's demand, and in so doing transform an ongoing lack of acknowledgment into the due regard we expect of one another.

*Conclusion*

My aim has been merely to spark a conversation about how beginning our theoretical investigations at the back end of blaming exchanges—with blame's normatively successful resolution via forgiveness—can produce genuine theoretical progress. To sum up:

1. Starting theoretical investigation at the front end of the blaming exchange yields indeterminacy about blame's function. If we start at the back end with normatively successful cases of forgiveness, however, we do get sufficient reason to believe its primary function is to demand. If the paradigm conditions for appropriately being forgiven consist most fundamentally in the taking up of a stance of remorseful empathic acknowledgment, and not the withdrawal of a false

claim about how the offended agent may be treated, then being appropriately forgiven is a matter of having acquiesced to a practical demand, not having repudiated a protested claim.

2. Leaning on resentment as the paradigm form of blame leads us mistakenly to believe that judgment is crucial to blame and (thus) forgiveness. When we start instead with forgiveness, whose normatively successful paradigm instances are conditioned on the (mere) empathic acknowledgment of the forgiven agent, we can see that judgment is superfluous to the process, and so resentment is not the relevant core of paradigm blame after all.

3. Agential anger's action tendency is to communicate itself to the offending party, and what it communicates is the demand for empathic acknowledgment. Such a demand makes most sense in light of the forgiven agent's previous failure of acknowledgment, a violation of the implicit default demand of daily interpersonal life. Forgiveness is thus fitting in response to the empathic acknowledgment of the offender in virtue of that acknowledgment having restored the normative equilibrium between the parties.

4. This account provides a more plausible resolution of the forgiveness puzzle than the judgmentalist resolution, given the latter's reliance on the theoretical heavy-lifting of judgment to resolve the puzzle, which is actually superfluous in the paradigm forgiveness—and blaming—exchange.

I realize that these are strong, revisionary conclusions. But I just aim to be starting a conversation, and so hopefully people will take these conclusions as what they really are, namely, provocations for more dialogue.[22]

[22] Acknowledgments (include members of Tulane Graduate Seminar on Agency and Blame, Spring 2015, as well as Per Milam and Massimo Renzo).

REFERENCES

Coates, D. Justin and Tognazzini, Neal, eds. 2013. *Blame: Its Nature and Norms*. Oxford: Oxford University Press.

D'Arms, Justin and Jacobson, Daniel. 2003. "The Significance of Recalcitrant Emotions (or, Anti-Quasijudgmentalism)." Reprinted in *Philosophy and the Emotions*, ed. Anthony Hatzimoysis. Cambridge: Cambridge University Press.

Darwall, Stephen. 2006. *The Second-Person Standpoint*. Cambridge, MA: Harvard University Press.

Davis, James R. and Gold, Gregg J. 2011. "An examination of emotional empathy, attributions of stability, and the link between perceived remorse and forgiveness." *Personality and Individual Differences* 50: 392-97.

Deigh, John. 1996. *The Sources of Moral Agency: Essays in Moral Psychology and Freudian Theory*. Cambridge: Cambridge University Press.

-----. 2011. "Reactive Attitudes Revisited." In Carla Bagnoli, ed., *Morality and the Emotions* (Oxford: Oxford University Press), pp. 197-216.

Dill, Brendan and Darwall, Stephen. 2014. "Moral Psychology as Accountability." In Justin D'Arms and Daniel Jacobson, eds., *Moral Psychology and Human Agency* (Oxford: Oxford University Press), pp. 40-83.

Franklin, Christopher Evan. "Valuing Blame." In Coates and Tognazzini 2013, pp. 207-223.

Frijda, Nico. 1986. *The Emotions*. Cambridge: Cambridge University Press.

Hieronymi, Pamela. 2001. "Articulating an Uncompromising Forgiveness." *Philosophy & Phenomenological Research* 62: 529-55.

Macnamara, Coleen. 2013. "Taking Demands Out of Blame." In Coates and Tognazzini 2013, pp. 141-161.

McGeer, Victoria. 2013. "Civilizing Blame." In Coates and Tognazzini 2013, pp. 162-88.

McKenna, Michael. 2012. *Conversation and Responsibility*. New York: Oxford University Press.

Scanlon, T.M. 2008. *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, MA: Belknap Press of Harvard University Press.

Scarantino, Andrea. 2014. "The Motivational Theory of Emotions." In D'Arms and Jacobson 2014, pp. 156-185.

Schmitt, M. et al. 2004. "Effects of objective and subjective account components on forgiving." *The Journal of Social Psychology* 144: 465-86.

Sher, George. 2006. *In Praise of Blame*. Oxford: Oxford University Press.

Shoemaker, David. 2014. "McKenna's Quality of Will." *Criminal Responsibility and Philosophy*. DOI: 10.1007/s11572-014-9322-5.

-----. 2015. *Responsibility from the Margins*. Oxford: Oxford University Press.

Smith, Angela. 2013. "Moral Blame and Moral Protest." In Coates and Tognazzini 2013, pp. 27-48.

Talbert, Matthew. 2012. "Moral Competence, Moral Blame, and Protest." *Journal of Ethics* 16: 89-109.

Thomas, Alan. 1999. "Remorse and Reparation: A Philosophical Analysis." In Murray Cox, ed., *Remorse and Reparation* (London, Philadelphia: Jessica Kingsley Publishers), pp. 127-134.

Walker, Margaret Urban. 2006. *Moral Repair: Reconstructing Moral Relations after Wrongdoing*. Cambridge: Cambridge University Press.

Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

-----. 2010. "Hypocrisy, Moral Address, and the Equal Standing of Persons." *Philosophy & Public Affairs* 38: 307-41.

-----. 2013. "Rightness and Responsibility." In Coates and Tognazzini 2013, pp. 224-223.

Watson, Gary. 2004. *Agency and Answerability*. Oxford: Oxford University Press.

Zechmeister J.S. et al. 2004. "Don't apologize unless you mean it: A laboratory investigation of forgiveness and retaliation." *Journal of Social and Clinical Psychology* 23: 532-64.

Jesse S. Summers
Duke University

**Bio:** Jesse S. Summers is a Post-Doctoral Fellow at Duke University. His research agenda focuses on understanding irrationality and its moral implications. His current research projects are on anxiety, especially whether anxious reasoning is like moral reasoning, addiction and compulsions, rationalization, and empirical challenges to philosophical accounts of agency. With Walter Sinnott-Armstrong, he is working on an extended project on Scrupulosity, a religious or morality-focused form of OCD.

# ***Post Hoc Ergo Propter Hoc*: Some Benefits of Rationalization**

Abstract: Research suggests that the explicit reasoning we offer to ourselves and to others is often rationalization, that we act instead on instincts, inclinations, stereotypes, emotions, neurobiology, habits, reactions, evolutionary pressures, unexamined principles, or justifications other than the ones we think we're acting on, then we tell a *post hoc* story to justify our actions. Although the conclusions of this research are in fact modest, I consider two benefits of rationalization, once we realize that rationalization is sincere. It allows us to work out, under practical pressure of rational consistency, which are good reasons to act on. Rationalization also prompts us to establish meaningful patterns out of merely permissible options.

Rationalization has a puzzling place in moral psychology: it is a profound challenge to our moral assessments of actions and agents, to whether we can praise and blame correctly, to our self-understanding and self-improvement, and to many of our metaethical views, particularly those that place prime importance on reasoning or deliberation, where rationalization occurs.[1] But it's a challenge that we largely ignore. Perhaps we assume that we generally know why we act, that our deliberation isn't systematically mistaken, and that any error or ignorance is easily enough remedied by some concerted introspection.

These assumptions are suspect, and the dangers of rationalization are serious.[2] One difficulty is that rationalization is most often sincere. But this sincerity also suggests some positive aspects of rationalization: rationalization puts practical pressure on us to work out good reasons for action and to assemble meaningful patterns where there were none. In this way,

[1] Simine Vazire and Erika N. Carlson, "Others Sometimes Know Us Better Than We Know Ourselves," *Current Directions in Psychological Science* 20, no. 2 (2011).
[2] Jesse S. Summers, "Rationalization and its Discontents," manuscript.

rationalization is an important instance of how explicit reasoning is morally relevant because it shapes and is shaped by our motivation.

## A General Account of Rationalization

Some psychologists have claimed that most explicit reasoning we offer to others or ourselves is rationalization, that we instead act on instincts, inclinations, stereotypes, emotions, neurobiology, habits, reactions, evolutionary pressures, unexamined principles, or justifications other than the ones we think we're acting on. Then we tell a post-hoc story to justify the actions that some underlying causes have already determined we'll do.[3] The challenge from this psychological data is that explicit, conscious reasoning is very often post-hoc rationalization.[4]

Rationalization requires there be at least two different explanations of the same action, one that is offered and another that is better. In a rationalization, the kind of explanation the rationalizer offers is a justification. The justification is cited *as if* it were an explanation of her action; however that particular justification is not a good explanation of the action. The better explanation may be a different justification, or, since not all explanations are justifications, the better explanation could be a habit, or a disposition, or an unconscious preference, or even a

[3] Note that this characterization of rationalization, here and throughout, covers two categories that should be distinguished for many purposes: motivated reasoning and confabulation. Motivated reasoning is explicit reasoning undertaken only under some psychological pressure to reach or avoid a particular conclusion. Confabulation is a justification or explanation offered when, in fact, one does not know why one acted. For only a small sample of the literature on the topic: Jonathan Haidt, "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment.," *Psychological review* 108, no. 4 (2001).; William Hirstein, *Confabulation: Views From Neuroscience, Psychiatry, Psychology and Philosophy* (Oxford University Press, 2009).; Benjamin Libet, "Do We Have Free Will?," *Journal of Consciousness Studies* 6, no. 8-9 (1999).; Daniel Wegner and Thalia Wheatley, "Apparent Mental Causation: Sources of the Experience of Will.," *American Psychologist* 54, no. 7 (1999).; Richard E. Nisbett and Timothy DeCamp Wilson, "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review* 84(1977).; Fiery Cushman and Joshua Greene, "The Philosopher in the Theater," in *The Social Psychology of Morality: Exploring the Causes of Good and Evil*, ed. Mario Mikulincer and Phillip R. Shaver (Washington, DC: APA Press, 2011).; Joshua D. Greene, "The Secret Joke of Kant's Soul," in *Volume 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, ed. Walter Sinnott-Armstrong (Cambridge, MA: MIT Press, 2007).; Michael S. Gazzaniga, *Who's in Charge?: Free Will and the Science of the Brain* (New York: HarperCollins, 2011). This distinction, while important for many purposes, should not matter to my discussion here.

[4] See Nisbett and Wilson, "Telling More Than We Can Know: Verbal Reports on Mental Processes." for an older survey of this and many other such examples, both of this form and of the next that I discuss. And, although the challenge is not about implicit reasoning, at least one line of response is to understand the underlying processes as implicit reasoning Terry Horgan and Mark Timmons, "Morphological Rationalism and the Psychology of Moral Judgment," *Ethical Theory and Moral Practice* 10, no. 3 (2007)..

biological, neurological, or other physical cause. In most cases, the rationalizer offers a worse explanation that she also *wants* to be the right explanation, but all that matters for rationalization is that the justification one offers as an explanation is worse than some alternative explanation.

The broad conclusion that our explicit reasoning plays only a minimal role in our actions is surely unwarranted, as there are clear counterexamples.[5] For example, we reason instrumentally about, say, how much money to withdraw from the ATM in order to pay a bill. But even if not all reasoning is rationalization, the research does show that we rationalize far more than sincere introspection reveals.

We cannot respond to this challenge by simply ignoring either the underlying causes of our actions or our explicit reasoning. We care not only why a person acts but also why she thinks she acts, and it would ignore large swathes of our ordinary action explanations to focus either on sincere justifications or on causal explanations. Therefore, I don't propose we respond to this challenge by simply asking which is one's "real" reason for acting: the underlying cause or the explicit justification.

If we are to take rationalization seriously, though, we won't get much guidance from philosophical literature, which has very little discussion of what rationalization is.[6] Nevertheless, I have previously argued that the following is a good general account of rationalization:

> A first-person rationalization, by *S*, of her *A*-ing, is a sincere, purported explanation of her *A*-ing, that she gives to herself or another, even after some introspection, which (a) offers a full or partial justification for her *A*-ing, (b) represents her *A*-ing as at least partially explained by this justification, but (c) some other justification or explanation better explains why *S A*-ed.[7]

This account makes clear that rationalization is not the same as lying, nor even of selectively sharing in a way that depends on my audience, which are both quite different

[5] The studies are often limited to circumstances in which reasoning would be less likely to be effective, and few attempt to explain away the obvious counterexamples, cases in which reasoning shapes our long-term plans, which we then execute. Alfred Mele, "Unconscious Decisions and Free Will," *Philosophical Psychology* 26, no. 6 (2013).; Darcia Narvaez, "The Social Intuitionist Model: Some Counter-Intuitions," in *Moral Psychology, Vol. 2: The Cognitive Science of Morality: Intuition and Diversity*, ed. Walter Sinnott-Armstrong (Cambridge, MA: MIT Press, 2008).

[6] Robert Audi, "Rationalization and Rationality," *Synthese* 65, no. 2 (1985)., 163.; Jason D'Cruz, "Rationalization as Performative Pretense," *Philosophical Psychology* 28, no. 7 (2015).

[7] Summers, "Rationalization and its Discontents."

challenges than rationalization.[8] Further, rationalization is not a simple failure of memory or reflection, so it must also be able to survive some genuine introspection by the rationalizer. (How much and what kind of introspection is hard to characterize in general terms.) Both of these caveats fall under the more general claim that rationalizations, unlike lies and simple mistakes, are sincere. I may "lie" to myself and "deceive" myself, but I do not do so knowingly if I'm rationalizing. Rationalizations are those justifications I do or would offer myself when I think about why I do what I do, which are not explicit attempts to fool myself or anyone else, but they still are not the best explanation of why I do what I do.

Of course, before moving on I have to acknowledge that our motives are often mixed, and most cases of rationalization are simplified here for the sake of discussion. In a real case, a rationalized justification *partially* explains an action, though one presents it as *fully* explaining the action. Or, the person may need *some* justification for his action, though any *particular* justification will be a rationalization. For example, I am motivated to buy a sports car because of mid-life doubts about my decreasing virility, but I'm responsible enough that I won't buy an expensive car without at least *some* justification for it. Therefore, the best explanation of my action is likely whatever causes me to search for that justification, though the justification has to be part of the story as well. Nevertheless, any *particular* justification I come up with—just think how much faster I can get to work in it!—is, taken by itself, only a superficial explanation of my action.

There is more to say about rationalization and its costs, but, given this account of rationalization, I want to show two of its benefits.

## Benefit of Rationalization: Consistency

Rationalizations are excellent—though not foolproof—ways of working out which justifications are good reasons. This is because rationalizations put us under some practical

[8] How I explain my actions to an interlocutor may change depending on the interlocutor, how close we are, how easily offended she is, etc. I never offer as my justification-as-explanation *everything* that I think went into my decision. Some of what into my decision not to go to dinner with my friend is too obvious to be worth saying (I did not believe he would kill me while at dinner), while other things may be true but not something my interlocutor would understand, or is more than I want to discuss with this particular interlocutor. The test here is what I say to myself, when sincere and even after introspection.

pressure to be consistent, which is pressure to determine what reasons we have that are consistent with past actions and to act in accordance with those reasons in future cases. I'll explain.

We feel pressure to act consistently. It would be easy to overstate this point: some people explicitly value spontaneity and unpredictability, and some circumstances are conducive to unpredictable behavior. But I mean that we feel pressure to act on a stable set of underlying motives, at least stable enough that we can make plans for our (future) selves. For example, if I find children overwhelmingly adorable on some days and disgustingly repellant on other days, I may have unstable preferences that make major life decisions difficult. But I hope that in fact my preferences and motives are stable. Perhaps I find dirty children repellant, or only toddlers, or perhaps I find only recently washed, smiling, quiet children adorable. In each case, stable preferences explain apparent conflicts, and I could use those stable preferences to help me make decisions for myself.

If we feel pressure to act consistently, where that means acting on consistent preferences or motives, yet we don't consistently know why we act, then how do we resolve this pressure?

First, we can speculate based on what makes the most sense of our actions and attitudes. I could guess that I really do like kids, but I don't like dirt, so that explains why I like clean kids but not dirty ones. It's a hypothesis I offer about my underlying motivation. I may have good hypotheses, but they are only ever that. When I rationalize, though, I sincerely claim that a particular justification was the explanation of my action. And this will apply that pressure to be consistent to our self-understanding and to our future behavior. Consider an extended example.

We are walking down the street together and someone in need asks me for money. We both know that there are many possible reasons for giving to someone in need who asks (and reasons against—but I'll ignore those here). The possible reasons to give are, for example, that it would alleviate some of his suffering, that god requires it, that it will put this man in his place, that it will impress you, and that I'm afraid of what he'll do if I refuse. Some of those are good reasons, some bad.

Let's say that I hand him money, and that the best explanation of my action is that I had intense fear when he asked, in part because of his race, and I gave over the money as quickly and unwillingly as if he'd mugged me. You then say to me, "I'm never sure what to do when people ask me for money: why did you give him money?"

As I'm rationalizing, what constrains the rationalization? I don't believe myself to be racist, so I don't consider that that could be an explanation. But my rationalization does have to be consistent with the action I'm rationalizing, i.e., I can't offer a justification that would not even *seem* to justify this particular action.[9] It must also be consistent with obvious past actions or motivations. Finally, it must be consistent with what I expect of myself in the future. How do these constraints look in this particular case?

I rationalize sincerely that I gave because he looked like he was suffering, and the money I gave would help alleviate that suffering. That's not the best explanation of my action, but it's the one I sincerely avow. My claim is not just that alleviating suffering *could* be a justification; my claim is that it *was* my justification, that this justification also explains my action.

In offering this rationalization, I imply my endorsement of the justification. (The implication is defeasible: "I gave to him in order to alleviate his suffering: what a fool I am when I forget the wisdom of Ayn Rand!") When I endorse the justification, this implies both that alleviating suffering *was* a good reason for giving to charity, and that alleviating suffering *is* a good reason to give to charity, at least in relevantly similar circumstances.[10]

Notice, now, what has happened and how this differs from a purely theoretical justification. There are many possible reasons to give to this man, and I may even think that some are better reasons than the one I offered. I might think, for example, that justice demands redistribution, and redistribution demands my giving money to this person, who needs it more than I do, and that this is a better justification than the one I (believe I) acted on. I might even wish that I'd been motivated by such an abstract motive as the need for redistribution. But I don't claim that I *was* motivated by redistribution. That would be a theoretical justification for my action, but not one that I endorsed by my action.

By claiming that I have actually endorsed this reason by my action, I now put practical pressure on myself—insofar as I care about or am committed to being a reasonable, consistent, and moral person who treats likes alike—to defend this as a good reason and act according to it

[9] In this case, my attempted justifications only has to be a plausible justification. If you ask me why I gave, and I say, "I love purple!", this is so far from a good justification (assuming the person wasn't dressed in purple) that it's hard to know whether to call it a bad justification or no justification at all. If, however, I just say, "I hate people who ask for money," then this doesn't *seem* to justify my action of giving money, but there still may be some connection, some suppressed premises, that I have yet to make clear.

[10] R.M. Hare, *The Language of Morals* (Oxford: Oxford University Press, 1952).; Thomas Nagel, *The Possibility of Altruism* (Princeton: Princeton University Press, 1970).

in future cases. If I turn the corner and walk past yet another person who appears similarly in need, I will feel some pressure to do one of the following: give to him as well, distinguish the cases ("this person doesn't actually seem to be suffering"), or add some nuance to the reason I previously endorsed ("…when I have extra cash in my wallet.")

Rationalization is then unlike a case in which I offer a theoretical justification. Imagine that I gave cash to the first person we walked by but I offered no rationalization of why I did so. Instead of offering you a reason, I say I'm not sure why I gave to that person when I don't normally give, and then you and I start a theoretical discussion about whether redistribution is a good reason to give to those in need. I say that it's a good reason, and we then turn the corner and see this new person in need. What kind of pressure do I feel? I feel no pressure to be consistent with my past reason: I offered no reason for my own actions. I may feel pressure to live up to those ideals that I just articulated, and that pressure can be significant, since I don't want to be a hypocrite. But I regularly fail to live up to my ideals: it's the price of maintaining high ideals. What I don't feel is pressure to be consistent with my past action, since I did not claim to have a justification that explained that past action.

Being consistent needn't necessarily lead me to being a better person: it may lead me to be a more *consistent* bigot, for example. But there is one way in which the pressure to be consistent may lead me to be a better person. If I tell you my (sincere) rationalization of why I gave to the man on the street—never mind that I was actually motivated by some racially motivated fear—you could challenge that reason. You could say that suffering leads to virtue and shouldn't be alleviated. In doing that, you don't just challenge my reason in the abstract. You're not just challenging me to come up with a better reason in an academic discussion. You're instead challenging something about *me*, about my values or my decision-making abilities or my perception of moral issues. I can't respond to that challenge by saying that redistribution is another possible justification, because I don't claim that it was my motivation in this particular case. It may be true that redistribution is a better justification, but your challenge was about what kind of things motivated me, not about how flexible I can be in coming up with new reasons when I'm challenged.

Further, if I ignore your challenge, that suggests more than just an intellectual flaw. It suggests that I am unreasonable or—depending on the particular reason—even immoral. When I offer a justification as an explanation, then I am liable to criticism in all these ways and have to

be prepared to stand behind the justification in a way that is not required when I offer a merely theoretical reason for acting in a particular way. Rationalization thus puts practical pressure on us to work out which considerations are good reasons. Perhaps that will even tend to lead us to be better people, though it should at least lead us to be more consistent people who are open to criticism.

Benefit of Rationalization: Meaningfulness

A second benefit of rationalization is the benefit of constructing meaningful explanations out of merely permissible patterns of action. I'll explain.

Most of my actions are obviously permissible. I walk into the cafe, order a drink, sit at an unoccupied table, turn on my laptop. Setting aside whether my first-world, energy-intensive, globally-exploitative, self-centered actions are, as a group, *all* impermissible, my quotidian actions raise no obvious issues of permissibility.[11] This is even true of my important decisions, like where to live or what career to choose, which are weighty and difficult but equally permissible.

Given that our actions are usually obviously permissible, we rarely consider moral reasons to prefer one action to another. In fact, in many cases, I may not consider non-moral reasons either. I may not ask myself where in the cafe (I ought) to sit or why I sat here rather than there. If I do consider possible justifications, they may not settle the issue: once I rule out the occupied and uncomfortable seats, the seats too far from outlets and too close to the

[11] I do take seriously the worry that many of my actions, in virtue of strong global inequalities, are impermissible. However, if most of my actions are impermissible because most of my options are impermissible (e.g., almost any food I can currently buy required objectionable exploitation to get it to within my grasp), then my choices are almost entirely moral blind alleys, where the morally best I can do is minimize my harm; analogously for the millions of people like me in such globally privileged positions, none of whom created these conditions. Unless there is an obligation to let myself starve, I still ought to choose certain options over others (e.g., buying fair-trade produce, though it also involves (less) objectionable exploitation), though all options are *ex hypothesi* impermissible. Saying that I ought to buy fair-trade coffee even if all imported food purchases are impermissible is perhaps not a contradiction, depending on how "ought" and "impermissible" are understood, but it's nevertheless a strange conclusion. It's particularly strange if the conclusion applies to all (or most) of the ordinary actions of entire populations—though that doesn't make the conclusion wrong. Cf. Immanuel Kant, *The Metaphysics of Morals,* trans. Mary Gregor (Cambridge: Cambridge University Press, 1996).), that the state may ensure that some of our actions, like buying property, are not systematically impermissible, though this may not apply to all actions, or to super-state problems.) But perhaps I'm merely rationalizing here, and I really just want my lifestyle to remain convenient.

restroom, I'm still left with several options. I "simply" pick one. I have no justification why I sat here rather than there, or why I set my coffee mug just here instead of slightly further away. That doesn't make my actions unthinking reflexes, but neither are they deliberated about or reflected on. They're habits, or "automatic," or dictated by social cues (smiling at someone smiling), or affordances (I grab the mug by the handle), etc.

And why care if we simply act for such causes and have nothing more to say by way of justification-as-explanation? Here is a real case. Someone asked me recently why I don't eat meat. I answered, honestly but sadly: "I don't know." This is not to say I am not serious about it. It's been decades since I first gave up meat, and I doubt I'll ever eat it again. I don't know for sure why I gave it up, but that initial cause hardly matters at this point. (Maybe it was what the cool kids were doing. Maybe it was to annoy my parents.) And I can certainly give some very good reasons why I shouldn't eat meat, or at least shouldn't eat much. But the only answer I am certain of as to why I don't eat meat isn't a justification at all, but is just that I don't really consider eating it. It doesn't occur to me to eat meat. I don't look at those parts of the menu, don't walk down those aisles in the grocery store, and don't consider it when someone offers it to me.

It is permissible and fairly easy not to eat meat, so I rarely think about a reason to do it. On the rare occasions it comes up, I can offer many justifications for not eating meat, any of which are sufficient for my not eating it. But there is now something lacking in my honest inability to offer a sincere justification-as-explanation. What is lacking is that I cannot distinguish—except by insisting—my pattern of action from one that is unintentional or held together by a mere whim. I would prefer to offer an explanation that would give this pattern of action some greater meaning, like an opposition to factory farming, ethical qualms about killing animals for food, or environmental concerns about using grain to make meat. I believe all of these are good reasons, and my choice not to eat meat would be a more meaningful pattern were I at least able to offer one of these justifications (to which I am committed) as in fact the reason I do not eat meat, as the justification that I can reasonably stand behind as explaining the pattern of actions.

This need for meaning arises especially when I have to choose among options that are vague, incalculably complicated, or relevantly incommensurable, so no option is clearly superior. For example, if I want to donate to charity but then have to choose which of many charities to

donate to, the problem isn't that they are similar in every relevant way, but that I don't know all of the relevant facts, and, even if I did, the facts still wouldn't settle for me what to do.

When I decide what to do and offer some justification as the explanation of my action, the justification is *compatible* with any actual motivation and with my action, even if it's not the only available justification. By committing to one particular justification, I do not think of my action as merely picking among the permissible. I have committed myself practically as caring about, say, early childhood poverty intervention as a reason to give to this charity over the others, thereby endorsing this as a reason, with all the entailments of this endorsement discussed above.

Of course, this doesn't explain why I feel pressure to justify any of my actions at all beyond their mere permissibility. Offering justifications beyond permissibility isn't obviously rationally or morally required—nor, for that matter, is introspection or even (much) self-awareness rationally or morally required.[12] Sometimes we do just pick. Regardless, rationalization, by offering a clear justification, is part of working out for oneself which reasons are salient and important, committing oneself to those reasons for the future, and thereby creating patterns of reasons out of merely permissible actions in a way that one may find meaningful.[13]

Moreover, once I see patterns in my actions, I may interact with this pattern more intentionally to shape future motivation. I may use the pattern to make predictions about myself. If I could believe that I don't eat meat because I care about animal welfare, this not only shapes the arguments I'm likely to make, but could shape the kind of discussions I have, the groups I join, the articles I read, the friends I make. What begins as a rationalization could transform into a genuine explanation of future actions, either an explanation of why I act to reinforce the pattern

[12] It can be easy to exaggerate this desire for justification, to think that we desire to justify all of our actions. Christine Korsgaard, *The Sources of Normativity* (Cambridge: Cambridge University Press, 1996).; T.M. Scanlon, "Contractualism and Utilitarianism," in *Utilitarianism and Beyond*, ed. Amartya Sen and Bernard Williams (Cambridge: Cambridge University Press, 1982).; J. David Velleman, "What Happens When Someone Acts?," *Mind* 101, no. 403 (1992).. I doubt this desire is present, even for the very reflective, except for those with certain anxiety disorders. Those disorders do lead a person to seek constant reassurance, which at least looks like a desire to justify most of one's actions. Nothing in non-pathological cases, though, looks like rational pressure to justify all of our actions.

[13] It's worth noting that this is not to claim that one creates "narratives" out of one's life, stories to tell ourselves as part of our self-understanding J. David Velleman, "Narrative Explanation," *The Philosophical Review* 112, no. 1 (2003).. Patterns are unnecessary for narratives, and narratives are unnecessary for patterns, though they're likely to co-occur.

or of how I try to reject the pattern. If I'd rationalized another explanation, things may have developed differently.

Rationalization therefore reveals that motivation is best understood not just as underlying motivations and explicit justifications. Our justifications can shape our underlying motivation over time in ways that are introspectively inaccessible. Deliberation does not always—perhaps does not usually—immediately issue in motivation. This is what the psychological work on rationalization may have right. Instead, explicit reasoning of the sort found in rationalization shapes one's motivation over time and is shaped by it, but not because deliberation controls motivation or vice versa. Rationalization, insofar as it requires explicit reasoning, can cover up why we act, but it can also be a crucial part of our ongoing process of changing our own motivation.

Notice what this discussion reveals about the most morally relevant case of rationalizing, and the one that some psychologists offer as a challenge to moral reasoning: moral dumbfounding.[14] There are certain moral claims that people hold, such as that incest is immoral. They are willing to offer reasoning as to why it is immoral: it's an improper violation of family trust; it leads to birth defects, etc. However, when a case is presented to a person in which each element of their reasoning is shown not to apply (the siblings are consenting adults, are conscientious about birth control, etc.) the person will be unwilling to give up their conclusion but also unable to offer any further reasoning to defend the conclusion. Moral dumbfounding is thought to show that our reasoning is therefore irrelevant to the moral positions we hold.

What I've suggested, however, is that our moral reasoning may only be irrelevant in the limited sense that it doesn't best explain why I am still motivated to make this particular judgment in this particular case. What is still true about my reasoning is that it is an attempt to make my judgments coherent and non-arbitrary: I endorse that violations of trust *are* good reasons to avoid certain relationships, and avoiding birth defects *is* a good reason to avoid certain pregnancy risks. (Notice that my reasoning could have been that violating biblical injunctions is a good reason, which would have been part of a very different pattern.) What moral dumbfounding points to is the pressure we feel to justify our judgments, how important our explicit reasoning is to us, and the discomfort we feel when the reasons we think support a

[14] Haidt, "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment."

judgment do not support that judgment. What it does not show is that our explicit reasoning plays no central role in our moral lives.

## Conclusion

I've argued that rationalization has at least two benefits. It puts practical pressure on us to work out good reasons for action and to assemble meaningful patterns where there were none. I've ignored here the costs of rationalization, so I cannot make a calculation of whether the benefits are worth those costs, but the fact that it has benefits at all is worth consideration.

I've ignored the most egregious cases of rationalizing, cases in which one rationalizes badly: a graduate student declares sincerely to her advisor that she is such a focused worker that she wouldn't notice if the house were burning down around her, sincerely thinking of the rare hours every few weeks when she is focused, ignoring the days in between during which she lacks focus and fails to work much at all. But even in these cases, which are the hardest to justify, there is some benefit. That student, biased as she may be to impress her advisor, and selective as she may be in her attention to the evidence, still puts *some* practical pressure on herself to become as focused as she imagines herself to be. And she may further come to think of those few periods of focused work as the ones that matter most to her (or that "define" her), which may have additional positive benefits, like keeping her from depressing thoughts about how badly she is doing.[15]

Therefore, even if rationalization is ubiquitous and one uses it to avoid tough truths about oneself or one's situation, rationalization *still* may have some benefit. Eliminating rationalization, even if it were possible, would reduce the positive effects that explicit reasoning has in our motivation more generally.

[15] The devil is in the details for a case like this, and it's hard to know if it would be good overall. An account as general as the one I develop here can't apply to specific cases without filling in a lot more than I can do here.

Zoë Johnson King

University of Michigan

**Bio:** Zoë is a PhD candidate at the University of Michigan. Before coming to Michigan, she did the BA and M.Phil in Philosophy at Cambridge, and then trained and worked for two years as a secondary school teacher in London. She has already informed too many people of her ludicrous dream to visit all 50 U.S. states by the end of her PhD to avoid actually doing it – tips are welcome.

# Trying is Good

## 1. The lay of the land

This paper argues that it is good to try to act rightly. By "good", I mean that trying to act rightly always goes some way toward making an agent a good person, with good character, and is something for which she deserves praise. By "trying to act rightly", I mean three things. First, being intrinsically motivated to perform actions that have the property RIGHT. Second, attempting to figure out what it is for actions to have this property, and thereby to identify right actions. Third, doing the things that you think are right, *because* they are right (or, at least, so you think!). Importantly, my view is that an agent *always* deserves praise for trying to act rightly; even if she fails, and ends up acting in a way that is either morally neutral or just plain wrong, the fact that she was at least *trying* to act rightly is still praiseworthy and still goes some way toward making her a good person.

The view that an agent deserves praise for trying to act rightly might strike the reader as obviously true. Indeed, it seems to be an empirical fact that we often do praise agents in this way, and that in some cases we do so based on no other considerations than that we think they deserve praise; locutions like "Well done, that was the right thing to do" or "Don't feel bad, you did what you thought was right" or "She is a very principled woman" are commonplace. But this view is in fact quite contentious. It is contested by Nomy Arpaly, in both her classic work, *Unprincipled*

*Virtue* (2003, pp.73-79), and her more recent book co-authored with Timothy Schroeder, *In Praise of Desire* (2013, pp.184-86). Other critics include Julia Markovits (2010) and Brian Weatherson (*ms*). These authors differ on the details, but what they have in common is roughly the following: they think that an agent deserves praise, not for trying *to act rightly*, but for trying to perform acts with whichever features turn out to be those that *make* acts right, according to the true first-order moral theory. Generally, these authors take no stand as to what the right-making features are. Their view is rather that, whatever they turn out to be, someone is good to the extent that she wants to perform actions with these features, and is praiseworthy to the extent that she is motivated to act rightly by her recognition that the right act has one or more of the features. On this view, whether an agent *wants* to act rightly – or *thinks* she is acting rightly – is irrelevant to her praiseworthiness.

Call this view the "Yoda view":

> **YODA VIEW**: for all features of actions *F*, if *F* is right-making, then an agent is praiseworthy to the extent that she is intrinsically motivated to perform actions with *F*. Trying to work out which features are right-making and to act accordingly earns no praise in and of itself.

The Yoda view is so-called because it can be glibly summarized in the aphorism "Try not. Do, or do not. There is no try".

Paulina Sliwa (2015) criticizes the Yoda view. Sliwa argues that someone is praiseworthy if and only if she wants to do the right thing, knows what the right thing to do is, and so acts rightly. Sliwa and I agree about the praiseworthiness of wanting to act rightly. But her position is more demanding than mine: she claims that trying to act rightly is *necessary* for praiseworthiness, while I claim only that it is *sufficient*. For example, Sliwa and the Yoda view would disagree about someone who is reliably moved by the suffering of others to alleviate that suffering, but who has no moral beliefs at all. Here I side with the Yoda view: withholding all praise from this agent is excessively harsh. Lousy abstract thinkers can be good people who do good things.

What I want to suggest is that the Yoda view is also too harsh. Intrinsic motivations whose objects are right-making features do contribute to praiseworthiness, but an intrinsic motivation *to act rightly* contributes to praiseworthiness too. It is no mark against an agent that the content of her motivation mentions the right-making features *de dicto* rather than *de re*; that her motivation is "I perform actions with the features that are right-making, whatever they may be" as opposed to "I perform actions with *F*". Performing actions with the features that are in fact right-making – in other words, performing right actions – is good. But it is not the only thing that is good. *Trying* to do this is also good.

## 2. Trying and succeeding

The importance of allowing praise for trying to act rightly is best shown through examples of agents who are not just *trying* to act rightly, but *succeeding* in doing so. My argument in this section will therefore proceed by discussion of cases. I will present three cases – one of an agent who is trying to act rightly and succeeding, and two of agents who manage to act rightly without trying. I will argue that there is no version of the Yoda view that can accommodate intuitively plausible verdicts about all three cases.

### *2.1 No harm in trying*

First, notice a consequence of the Yoda view. The view entails that intrinsic motivations to act rightly do not contribute to praiseworthiness (as I think they do), no matter what the true first-order moral theory turns out to be. This is because, whichever features of actions turn out to be right-making, rightness itself cannot be among them; actions cannot be right *in virtue of their rightness*. That would be circular. The right-making features must be something else. Moreover, the Yoda view also entails that no non-intrinsic motivations contribute to praiseworthiness. This follows directly from the view – the view explicitly states that only intrinsic motivations count. Putting these two entailments together highlights an important consequence of the Yoda view. An agent who is intrinsically motivated to act rightly, and who wants to perform actions with the

features that are in fact right-making only *because* she believes (correctly!) that possessing these features constitutes an act's rightness, is *not at all praiseworthy* according to the Yoda view. Such an agent would indeed be motivated by the right-making features. But the relevant motivations are not intrinsic; in Arpaly and Schroeder's terminology (2013, pp.6-9), they are "realizer" desires. And this agent's only intrinsic motivation is directed toward rightness itself – a feature of actions that cannot be right-*making*. So this agent does not have any motivations that are both intrinsic and directed toward features of actions that are right-making. Thus, there is nothing to render her praiseworthy.

Now consider the following case:

> ***Burgers:*** In College, Gottlob couldn't wait to start taking Ethics classes, to clarify his thoughts about his obligations toward refugees, the environment, non-human animals, and so on. After years of careful study, long conversations, and much consternation, he eventually came to accept a form of pluralist consequentialism as the true first-order ethical theory. And Gottlob revised his behavior accordingly. For example, he buys only organic meat, because he thinks that the comparative value of human gustatory pleasure and the avoidance of harm to animals and the environment – coupled with his views about our obligations in collective action problems – entail that buying regular meat is impermissible, but buying organic meat is permissible. Moreover, *Gottlob is completely right about all of this*. Not only is pluralist consequentialism the true first-order moral theory, but the list of values that Gottlob endorses corresponds *exactly* to what really is valuable, to the precise comparative degrees of value that these various things really have. Gottlob has totally nailed Ethics. And, since he is not at all akratic, he also acts perfectly.

What should we say about Gottlob? He is trying to act rightly, and he is doing a great job. He acts impeccably. Gottlob also has all and only the true moral beliefs. And this is no fluke; it is due to years of careful thought and sophisticated reasoning – in short, due to a great deal of effort on his part. Yet the Yoda view entails that Gottlob is *not at all praiseworthy*. This is because,

though he is motivated by all the right-making features of actions to the precise degrees given by the true moral theory, he is not so motivated *intrinsically*. He is motivated by these features of actions only because he believes that they are right-making, making the desire that underpins this motivation a "realizer" rather than an intrinsic desire. If Gottlob were to change his beliefs about what it is that rightness consists in, his motivations would likewise change. His only intrinsic motivation is to act rightly. But rightness itself is not right-*making*. Thus, there is nothing to render Gottlob praiseworthy.

Contrast ***Burgers*** with this case:

> ***Burgers 2***: Friedrich has always found himself with certain intrinsic motivations, directed toward various objects. It is because of this that he came to adopt a certain kind of pluralist consequentialism back in College, having recognized that this theory best reflects his pre-theoretical intuitions. But his commitment to the theory is an idle wheel. Friedrich does not actually care about his actions' moral status. So he could be convinced of a different view of what rightness consists in without changing his motivational structure one iota. He will continue, for example, to buy organic burgers rather than regular burgers, since he is intrinsically motivated by gustatory pleasure to *some* extent, but this is outweighed by his intrinsic aversion to harming animals and the environment when he thinks about buying burgers.

The Yoda view said that Gottlob was not at all praiseworthy. But the Yoda view says that Friedrich is *fully* praiseworthy. After all, Friedrich's motivations perfectly match the content of the true moral theory, *and* he is so motivated *intrinsically*!

I want to suggest that this is clearly the wrong result. Friedrich's moral status cannot be so far removed from Gottlob's. After all, we have stipulated that their motivations and the moral facts are such that both agents act impeccably; both of their lives consist in the performance of right action after right action. And both of these agents have all and only true moral beliefs. (We can even stipulate that they are also both correct about all relevant non-moral matters.) They both

believe that their actions are right, that they are value-maximizing, and that they have the former property in virtue of having the latter. They completely agree about what is valuable. Indeed, Gottlob and Friedrich diverge only in what we may most accurately say about the description "under which" they are motivated to act: Gottlob buys his organic burgers on the grounds that doing so is *right*, whereas Friedrich buys them on the grounds that doing so is *value-maximizing*. But since both agents believe – truly, we assume – that maximizing an appropriately weighted combination of the values on their mutual list is what constitutes rightness, they both think that the properties "value-maximizing" and "right" are at least co-extensional, and possibly (if moral facts are metaphysically necessary, as is often assumed) co-intensional. So the agents themselves do not see this difference between them as terribly important.

To repeat: the Yoda view entails that, while Friedrich is fully praiseworthy, Gottlob is *not at all praiseworthy*. This is an extremely unwelcome conclusion, in light of the observation that Gottlob is not only extremely similar to Friedrich, but is pretty much a moral saint. So I submit that the Yoda view must be rejected.

*2.2 An easy fix?*

Could there be a modified Yoda view that stops short of conceding that motivations to act rightly are themselves praiseworthy, while giving an intuitively plausible verdict about ***Burgers***? There could. The Yoda view entails that Gottleb is not at all praiseworthy because it restricts our focus to *intrinsic* motivations. So an easy fix would be to allow that "realizer" motivations – i.e. motivations to φ that develop because the agent is already motivated to ψ and comes to believe that φ-ing constitutes ψ-ing – also contribute to an agent's praiseworthiness. (This is contrary to Arpaly and Schroeder's views, but we can deviate from them.) And if we decide that realizer motivations also contribute to praiseworthiness, then we have no problem giving the intuitively correct verdict about Gottlob. Granted, he does not have any *intrinsic* motivations for that which is right-making. But he has a full set of the corresponding realizer motivations.

Call this view the "new Yoda view":

> **NEW YODA VIEW**: for all features of actions *F*, if *F* is right-making, then an agent is praiseworthy to the extent that she has either an intrinsic or a realizer motivation to perform actions with *F*.

The new Yoda view gets the intuitively right result about Gottlob. But, in so doing, it opens itself up to an absurdly large number of new counterexamples. While the Yoda view is too harsh, the new Yoda view is not harsh enough.

Here is an example:

> ***Aardvark approval***: Aarulina desperately wants to act in a way that is approved of by aardvarks. This serves no further end; just as some people are intrinsically motivated to act in a way that is approved of by their friends and loved ones, or by God, for Aarulina it's all about aardvarks. But, fortunately, Aarulina thinks that she has figured out what it is that aardvarks approve of, and therefore what it is to act in a way that's approved of by aardvarks. According to Aarulina, aardvarks approve when people perform actions with feature *F*.

Examples like this are easy to invent: ascribe to your fictional agent a bizarre but morally neutral intrinsic desire, the (plainly silly) belief that performing actions with feature *F* constitutes achieving the object of their intrinsic desire, and the corresponding realizer motivation to perform actions with feature *F*. But if feature *F* is, in fact, right-making,[1] the new Yoda view is in trouble. This is because the new Yoda view entails that oddballs like Aarulina are *fully praiseworthy*, just like Gottlob. But these people are crazy people, not moral saints. Gottlob surely must be more praiseworthy than these agents. After all, he has a realizer motivation to perform actions with the right-making feature because he is trying to act rightly and has succeeded fantastically, but they have realizer motivations to perform actions with a right-making feature as a weird coincidence. Any theory that entails that Gottlob is no more praiseworthy than these crazy people must be rejected. So much for the new Yoda view.

[1] Since I have assumed nothing about *F*-ness, it should not be difficult to stipulate that it is right-making. If the true moral theory says that several different features are each right-making, then let *F* be their disjunction.

This amounts to a dilemma. If the Yoda view allows realizer motivations to contribute to praiseworthiness, then it gives intuitively incorrect verdicts about cases like ***Aardvark approval***. But if it doesn't, then it gives intuitively incorrect verdicts about ***Burgers*** and ***Burgers 2***.

A better response to cases of people trying to act rightly and succeeding is to accept that these people are praiseworthy. We should say that we cannot earn praiseworthiness by deriving realizer motivations for that which is right-making from just *any* old intrinsic motivation; that motivation must itself be praiseworthy. But this poses no problem for people like Gottlob, since being motivated to act rightly is indeed praiseworthy. Trying is good.

## 3. Trying and failing

I have provided support for my view, and undermined the Yoda view, based on a case of an agent who is trying to act rightly and succeeding. But defenders of the Yoda view sometimes think that their view is supported by cases of agents who are trying to act rightly but failing. Here is an example from Arpaly and Schroeder (2013, pp.183-4):

> Consider a person who keeps slaves because he takes it to be right… we hold that the fact that he believes having slaves to be [right] is no excuse for his actions.

And here is one from Markovits (2010, p.224):

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

Let's think about whether our intuitions about agents who try to act rightly but fail really do lend support to the Yoda view.

When an agent is "in the grip of a false moral view" – to borrow a phrase from Harman (2015) – and she wants to act rightly, she may end up acting wrongly. For example, if she sincerely believes that she is morally required to refuse to conduct same-sex marriages, when in fact this is morally forbidden, then her trying to act rightly will lead to her acting wrongly. Harman (2011, 2015, *ms*) argues, as do Arpaly and Schroeder and Markovits, that such an agent is blameworthy for having acted wrongly, and is not "exculpated" by her false moral beliefs.

This might seem like a challenge for my view. But motivations, actions, and beliefs are all quite different things. Someone can be blameworthy for one or two of these things without being blameworthy for all three. We can blame her for acting wrongly, or for having false moral beliefs, or both, without piling extra condemnation of her motivations on top. This is what I suggest we do when we consider agents who try to act rightly but fail.

In support of my proposal, note that agents can be led astray by false moral beliefs even when their motivation is not *to act rightly*, but *to perform actions with F,* where *F* is in fact right-making. For example, suppose that fairness is a right-making feature, and consider this case:

> ***Fairness***: A father wants to decide on a toy-sharing policy for his daughters. He is committed to the value of acting fairly, and wants his toy policy to be fair. So he thinks awhile and develops a rudimentary theory of fairness. But he gets it wrong: he thinks that his girls' age-difference is irrelevant to considerations of fairness, when in fact it is highly relevant. So he ends up choosing a policy that is unfair to his younger daughter, and therefore wrong.

The Yoda view says that the father is praiseworthy, since he is intrinsically motivated by a feature of actions that is, in fact, right-making – fairness. But this agent is still led to act wrongly by his false moral beliefs. Again, it is easy to construct examples like this: describe your fictional agent so that, rather than accepting a false theory about what the right-making features *are*, she accepts a false theory about what one of the (genuine) right-making features *itself* consists in. This is still a false moral theory; theories of fairness, well-being, promise-keeping,

and so on are moral theories.[2] And false theories on these topics lead people like the father to act wrongly. So if the defender of the Yoda view wishes to say, with Harman, that false moral beliefs do not exculpate, then she must say that these agents too are blameworthy for their wrongful actions.

I do not think that there is any tension here. The Yoda view can accept Harman's account of blameworthiness for actions. We can say that the father is still praiseworthy insofar as he was *trying* to act fairly, though he remains blameworthy insofar as he *in fact* acted unfairly. We can say, "He messed up, but at least he had good intentions".

But if we are going to say this about ***Fairness***, we might as well say it about agents who try to act *rightly* but fail to do so. Once we allow for an agent to act wrongly (and so be somewhat blameworthy) based on false moral beliefs (for which she may also be blameworthy), but still to have had good intentions (and so be somewhat praiseworthy), we can extend this analysis from agents whose intention was to perform actions that instantiate some right-making feature to those whose intention was to act rightly. Indeed, it is not at all obvious why the moral status of trying and failing to act, say, *fairly* should be so different from that of trying and failing to act *rightly*, if in both cases the agent fails due to her false moral beliefs. So the defender of the Yoda view at least owes us an account of this difference. Without such an account, our intuitions about agents who try to act rightly but fail do not really lend support to the Yoda view.

4. Not trying, but succeeding anyway

There is one final type of case that I should address when discussing the Yoda view. Those who defend this view typically employ two types of case: one in which an agent has false

[2] One exception is for right-making features that can be specified in fully non-moral terms, if there are any of these. If a right-making feature can be specified in non-moral terms, then false beliefs about which actions have this feature will be false *non*-moral beliefs, rather than false moral beliefs. And this might be an important difference (see, e.g., Harman 2015 pp.62-69). However, this point should be of little comfort to defenders of the Yoda view who think that the plausibility of their view does not depend on what the right-making features are. If they have to assume that all right-making features can be given a fully reductive non-moral analysis, then the tenability of the Yoda view clearly is dependent on which first-order moral theory turns out to be true.

beliefs about what is right, tries to act rightly, and ends up acting wrongly (as I have already discussed), and one in which an agent has false beliefs about what is right, but is nonetheless moved by what really *are* right-making features, and ends up acting rightly after all. For this second type of case, these philosophers usually concentrate on the example of Huckleberry Finn, who is moved by the importance of freedom and equality despite his officially not believing in it (see Arpaly 2003 pp.9-10, Arpaly and Schroeder 2013 p.178, Markovits 2010 pp.208-209, Weatherson *ms* pp.67-68). Their verdict is that Huckleberry is praiseworthy for being motivated to act rightly by the right-making features, notwithstanding the fact that he falsely believes that his actions are wrong.

But I am comfortable with this verdict. I am happy to say that being motivated by right-making features is one way to be good. The claim that I am defending in this paper is that it is not the *only* way to be good; trying to act rightly is also good. This means that cases of agents who don't try to act rightly, but still succeed in acting rightly because they are motivated by the real right-making features, are uninteresting test cases in the present dispute – no cases like this tell *against* the view that agents are *also* praiseworthy for trying to act rightly.

## 5. Conclusion

I have argued that it is good to try to act rightly. I have suggested that we must embrace this view to avoid implausibly harsh verdicts about agents who try to act rightly and succeed, while my view has no more trouble than its opponents in accommodating our intuitions about agents who try to act rightly but fail.

Perhaps I have failed to convince the reader. But at least I tried.

**REFERENCES**

Arpaly, Nomy (2003). *Unprincipled Virtue*. Oxford: Oxford University Press.

Arpaly, Nomy and Timothy Schroeder (2013). *In Praise of Desire*. Oxford: Oxford University Press.

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

Harman, Elizabeth (2015). "The Irrelevance of Moral Uncertainty". *Oxford Studies in Metaethics* 10, 53-79.

Harman, Elizabeth (*ms*). "Ethics is Hard! What Follows?"

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

Markovits, Julia (2012). "Saints, Heroes, Sages and Villains", *Philosophical Studies* 158, 289-311.

Sliwa, Paulina (2015). "Moral Worth and Moral Knowledge", *Philosophy and Phenomenological Research*, first published online June 2015, DOI: 10.1111/phpr.12195.

Weatherson, Brian (*ms*). *Normative Externalism.*

Monique Wonderly

Princeton University

**Bio:** Monique Wonderly is the Harold T. Shapiro Postdoctoral Research Associate in Bioethics at the Princeton University Center for Human Values. Her primary research areas include theoretical and applied ethics, moral psychology, and the philosophy of emotion. Her current research focuses on emotional attachment – and in particular, on questions concerning moral agency and ethical treatment that arise when considering certain attachment-related pathologies, including psychopathy and (some forms of) addiction.

# The Value of Attachment[1]

There is a rich tradition in ancient thought that instructs against attachment to others.[2] Historically, Stoics, Buddhists, and Daoists have denounced attachment on the grounds that it renders us vulnerable to suffering and/or interferes with autonomous agency. Contemporary Western philosophers have often dismissed such views as failing to adequately recognize the import of caring relationships. Caring, despite its potential to cause suffering and undermine autonomy, is thought to have an immense value that typically outweighs its costs. But even if one grants that views that instruct against *caring* are non-starters, this doesn't settle the matter. Attachment is not synonymous with caring. Once we disentangle these attitudes, we can no longer simply assume attachment's value based on that of caring. Here, I argue that attachment – even when not construed as a kind of caring – can have great value for an agent insofar as it constitutes a rich form of needing another.[3]

In what follows, I begin by articulating a conception of attachment on which it is distinct from caring. Next, I review various worries associated with felt necessity and suggest how they might be mitigated in the case of caring but not necessarily in attachment. Finally, I argue that

[1] Many thanks to Coleen Macnamara, Agnieszka Jaworska, David Beglin, Ruth Chang, Elinor Mason, Eric Schwitzgebel, John Martin Fischer, Dana Nelkin, Andrews Reath, Maudemarie Clark and Luc Bovens for instructive critical comments on various aspects of this work.

[2] On my view, we can be attached to non-person, and even to ideas, but I will largely restrict discussion here to attachments to other persons. Both the terms "attachment object" and "attachment figure" will refer to a person to whom one is attached.

[3] I say "can" because I take it that attachment, like most attitudes, is not unconditionally valuable. Some forms of attachment – e.g., malicious or exploitative forms, may be on the whole disvaluable.

experiencing another person as a felt need *qua* attachment can be a rich source of value, and I explore how the preceding discussion points to an interesting, complementary relationship between caring and attachment.

## §1. Attachment and Caring [4]

In a previous work, drawing on views from ancient Stoicism, Eastern philosophy, and psychology, I articulated the key marks of a particular kind of attachment that is distinct from caring. In attachments of the relevant sort, the attached agent (I) has a relatively enduring desire for engagement with a non-substitutable particular, (II) tends to suffer a reduced sense of security upon prolonged separation from the object, and (III) tends to experience an increased sense of security upon obtaining the desired engagement with her attachment object. Importantly, *security*, as I use the term, does not denote mere feelings of "safety" or "comfort." Rather, security is construed as a kind of confidence in one's well-being and agential competence. In colloquial terms, upon prolonged separation from our attachment objects, we often feel "off-kilter," "no longer all of a piece," "as though we've lost our bearings," etc. Conversely, engagement with our attachment objects allows us to feel as though we are "on solid ground," more together, and more competent.[5]

I refer to attachments of this sort as security-based attachments.[6] A paradigmatic case of security-based attachment is the infant-primary caregiver bond. Let's start with an example.

> Tommy is a twenty month-old who is attached to his mother, Marie. Though he has a large family to care for him, Tommy tends to seek out Marie specifically for cuddles and play. He becomes anxious when she's away for too long, and if she's not around when he is injured or frightened, he often becomes inconsolable. When Marie is nearby, Tommy is more willing to try new activities and to engage with new playmates.

Tommy's behavioral pattern toward Marie represents a typical infant-primary caregiver attachment. According to John Bowlby, known as the father of "attachment theory," the infant-

[4] Some passages in this section appeared in Wonderly 2016.
[5] This conception is largely drawn from the psychological literature on security and attachment. See, for example, Maslow (1942); Blatz (1966); Bowlby (1969/1980); and Ainsworth (1988).
[6] In the remainder of this paper, any use of the term "attachment" refers to security-based attachment.

primary caregiver bond is characterized by a set of evolutionarily adaptive behaviors that serve to provide the infant with a sense of security. The attached infant attempts to remain in close proximity to her primary caregiver, treats her as a ‘‘secure base’’ from which to explore unfamiliar surroundings, seeks her out for protection as a ‘‘safe haven’’ when threatened, and protests separation from her via crying and other displays of distress.[7]

Interestingly, psychologists have noted that versions of these behaviors are also typically present in long-term adult romantic partnerships. Consider another example.

> Adam and Linda have been married for twenty years. Adam regards Linda as his rock, turns to her first when he is troubled, and feels more confident and capable when she is nearby. With her beside him, he is more comfortable taking on risks and new challenges. When the pair are separated for prolonged periods, even while spending time among friends, he tends to feel a bit “off” and can’t seem to get along quite as well as usual.

This case represents a rather typical, long-term romantic attachment. As psychologists have noted, adults do tend to seek proximity to their romantic partners and protest long-term separation from them. Our romantic partners also function both as secure bases and safe havens for us. When our romantic partners are nearby, we feel more competent to explore new environments and to take on challenging situations. Also, we tend to turn specifically to our romantic partners for comfort and support during periods of significant stress.[8]

On my view, we can become attached to a variety of persons (and objects), and I focus on the *affective* orientation of attachment relationships.[9] Importantly, to be attached to a person in the relevant way is to experience her as a felt need, such that without her, one tends to suffer a reduced sense of confidence in one’s well-being and agential competence. Despite obvious differences in the two cases, Tommy and Adam are attached to Marie and Linda, respectively.[10]

[7] Bowlby 1969/1980

[8] See for example Rholes & Simpson (2004); Brumbaugh & Farley (2006); Hazan, Campa, & Gur-Yaish (2006); Collins et al (2006) and Mikulincer & Shaver (2007).

[9] See Wonderly (2016) for a more thorough account of my view of attachment and how it compares to traditional views of attachment theory in developmental and clinical psychology.

[10] One relevant difference is that Adam has a developed sense of self that can be upset in more complex ways, and on account of that, his reduced sense of security will likely consist in a richer affective experience. Relatedly, one might wonder whether toddlers can have a sense of security in the sense I described above. Confidence in how one is faring and in how well one is able to navigate the world, for example, would seem to require having some idea of one’s own well-being and one’s own agential competence – concepts which are likely inaccessible to the average twenty-month old. Though toddlers are not capable of *reflecting* on their own well-being or their own agential

Though it is often the case that we care about our attachment figures, attachment isn't synonymous with caring. Theorists generally describe caring in terms of an emotional vulnerability to how the cared-for object is faring and certain desires to promote its flourishing.[11] Presumably, Adam is not only attached to, but also cares about, Linda. And insofar as he does, he will be disposed to promote her good and to experience emotions that track ups and downs in her well-being (sadness when she fares poorly, happiness when she thrives, etc.).

Importantly, one can care about another without being attached to her. Adult siblings and friends who live far apart often have this type of relation. Suppose, for example, that Adam cares about his sister Sara, but the two live on opposite sides of the country and their busy schedules preclude frequent interaction. In conversations with other relatives, Adam often asks how Sara is doing and how he might help her do better. He feels joyous when she's faring well, upset when she is struggling, relieved when she prevails, etc. All the same, he is content to see her only on special occasions, getting along just fine without regular engagement during the in-between periods. In other words, engagement with her – or the lack of it – doesn't impact Adam's sense of security.

One can also be attached to another without caring about her. Let's return to Tommy's attachment to his mother. One reason to doubt that Tommy cares about Marie is that, on many accounts, he doesn't yet have the cognitive capacities required for caring.[12] More importantly, though, Tommy's affective orientation toward Marie isn't obviously focused on her well-being. Recall that attachment, unlike caring, is not in the first instance about the flourishing of its object, but about the way in which engagement with a non-substitutable particular impacts one's own sense of security. In this way, attachment is more self-focused than caring tends to be.

Adults can have non-caring attachments as well. Imagine, for example, that Adam has a non-caring attachment toward his personal trainer, Trent. Suppose that he has a strong desire to receive work-out instruction from Trent specifically – no substitute will do, that he becomes unsettled and mildly distressed when Trent is unavailable, and that he feels more "empowered"

competence, they certainly can experience affects that track these features. For example, they can feel confused, anxious, frustrated, and reluctant to do, or unable to do, the things that they normally can.

[11] See, for example, Frankfurt (1999b); Shoemaker (2003); Jaworska (2007a, 2007b); Helm (2010); and Seidman (2008). While Frankfurt (1999a/b) specifically stresses the *non-emotive* features of caring, in an earlier work, he does tie the notion of vulnerability to caring (1988).

[12] Caring theorists have suggested that caring requires the abilities to have higher-order desires (Frankfurt 1999b, p. 161), have the concept of importance (Jaworska 2007a, p. 561), and/or see the object of one's care as a source of reasons (Seidman 2008, p. 12).

when working out with Trent in particular. Though Adam ought to recognize that Trent is a person who has ends all his own and deserves to be treated accordingly, his thoughts and feelings needn't be tethered to Trent's well-being in the robust sense that caring (as described above) requires. For instance, it might be reasonable for Adam to remain relatively unbothered by mishaps that negatively impact Trent's life (assuming they don't affect his performance as a trainer), where we wouldn't expect such apathy from someone who genuinely cared about him.

In sum, attachment, unlike caring, necessarily involves experiencing another as a felt need and an essential connection between that individual and the attached agent's sense of security. Given that these attitudes are distinct, one cannot directly infer the value of attachment from the value of caring. I will argue that attachment has value in virtue of the type of felt necessity internal to it, but first let's consider the phenomenon of felt necessity more broadly.

## §2. Felt Necessity

Roughly, to need something is to be such that one would be, in some sense, harmed without it.[13] Whether or not one actually needs someone or something, one might nevertheless experience that person or object as a felt need. The phenomenological character of this experience seems to differ from mere desire. There are things that we want, and then there are things that we feel that we *must have*. We often use the language of need in everyday discourse to mark this distinction. Consider the familiar utterance: "I don't simply want it, but I *need* it."[14]

Intuitively, having needs – or at least acquiring new ones – seems rather unfortunate. To need another, after all, is to be subject to suffering and dependence. [15] Vulnerability to suffering and dependency underlie some of the strongest criticisms of attachment that we find in ancient Stoicism and Eastern philosophy, and while both qualities are pervasive aspects of human life, they continue to be viewed as largely negative.[16] In this section, I briefly explore these worries and juxtapose the type of felt necessity in caring with that of attachment.

[13] See for example Harry Frankfurt (1999a) and David Wiggins (1998).
[14] Not much hinges on this distinction. One might think that felt necessity is just a strong type of desire, but the point here is just that needing, as opposed to merely wanting, is often characterized by a sense of urgency and motivational priority.
[15] For examples of theorists who have explicitly associated needing with one or both of these features, see Miller (2012), Wollheim (1973), Stampe (1988), Wiggins & Dermen (1987), and Frankfurt (1988). Alasdair MacIntyre (1999) and Erinn Gilson (2011) both acknowledge (and reject) the common view that vulnerability and dependence are essentially negative.
[16] See, for example, Cicero (1887); Epictetus (1983), 3.24; *Digha Nikaya* (1987: 31, 34); and the *Chuang Tzu*: *The Inner Chapters* (1981 Graham, trans.).

Susceptibility to harm is a ubiquitous feature of all creatures, but nevertheless, on most accounts, a quite unfortunate one. We typically seek to *mitigate* vulnerability by becoming stronger or implementing defensive measures. Placing high value on the avoidance of harm or suffering, despite its rather myopic focus in some Stoic and Asian philosophies, is not restricted to ancient or Eastern views. Nearly all contemporary Western theories of well-being name suffering as a key mark of ill-being.[17] The point, of course, isn't that we should avoid suffering *at all costs*, but rather if an object or state of affairs entails – or is likely to bring about – suffering, then that fact alone counts as a mark against it.

Dependence, like vulnerability, is also often considered a ubiquitous but regrettable aspect of human life. Being dependent upon others, at least after childhood, is associated with *weakness*. We often consider it a good thing when an agent becomes more "autonomous" by reducing her dependence on others. Here again, while the ancient Stoic emphasis on self-sufficiency has lost some traction in the modern era, views that laud independence as a vital aspect of autonomy are in no short supply.[18] Though some degree of dependence on others is inevitable, it makes sense to think that *ceteris paribus* it is better to decrease our dependence on others overall.

In virtue of rendering us vulnerable and dependent, experiencing some external person or object as a felt necessity tends to constrain one's agency in certain ways. Felt needs generally have a kind of motivational priority, and this makes them potentially dangerous. They tend to exert a more demanding influence over our thoughts, feelings, and actions than typical desires. Our needs can capture and fix our attention, and when they become pressing enough, they often diminish the salience of other of our important concerns and undermine self-control.

For a vivid if admittedly extreme example of this phenomenon, consider novelist and self-described heroin addict, William S. Burroughs, on heroin's "algebra of need": "A dope fiend is a man in total need of dope. Beyond a certain frequency, need knows absolutely no limit or control. In the words of total need: "Wouldn't you?" Yes you would. You would lie, cheat, inform on your friends, steal, do anything to satisfy total need…"[19] Of course, felt necessity is

[17] See Bentham (2007); Griffin (1986); Kraut (2009); and Crisp (2013).
[18] For a thoughtful discussion and critique of such views, see Govier (1993, pp. 100-104).
[19] Burroughs 1959, p. xxxvii. Consider also another excerpt from the same work: "I was only roused to action when the hourglass of junk ran out. If a friend came to visit…I sat there not caring that they had entered my field of vision…and not caring when he walked out of it. If he had died on the spot, I would have sat there looking at my shoe waiting to go through his pockets" (ibid, xiii).

rarely so pernicious, but as Gary Watson notes, even the most cherished varieties of felt necessity can constrain one's agency in ways comparable to that of addiction. Watson explains, "Like addictions, to be attached in [ways exemplified by caring relationships such as parenting or romantic love] is to be vulnerable to diminished control of certain kinds."[20]

Felt necessity, then, on account of rendering us vulnerable to suffering and dependent on others, appears to be a potentially onerous condition. Yet, it is one that we readily countenance in certain kinds of caring relationships – presumably because caring has such value otherwise. To see whether the value of caring can help to illuminate the value of attachment, let's take a closer look at the type of felt necessity involved in caring (henceforth, caring necessity). Here, it will be helpful to review Harry Frankfurt's work on the necessities of love.

In "On Caring," Harry Frankfurt explains that in virtue of loving some person or object, we are typically subject to a kind of volitional constraint that is experienced as felt necessity. We feel as though there are certain things that we *must* do – or again, *mustn't* do – in virtue of our love. Frankfurt writes, "It is characteristic of our experience of loving that when we love something, there are certain things that we feel we *must* do. Love demands of us that we support and advance the well-being of our beloved, as circumstances make it possible and appropriate for us to do so; and it forbids us to injure our beloved, or to neglect its interests."[21]

On Frankfurt's view, the need to do (or to avoid doing) certain things in service of one's love is tightly connected to caring necessity, i.e., the need for one's beloved to flourish. He explains, "…the well-being of what a person loves is for him an irreplaceable *necessity.* In other words, the fact that a person has come to love something entails that the satisfaction of his concern for the flourishing of that particular thing is something that he has come to need."[22] On Frankfurt's account, love is a mode of caring, and in caring about another, an agent becomes "vulnerable to losses and susceptible to benefits depending upon whether what he cares about is diminished or enhanced."[23] Caring theorists also generally agree that this orientation toward

[20] Watson 2004, p. 85. Watson is among the agency theorists who explicitly associate addiction with felt necessity. He writes, "To become addicted…is to acquire a felt need, a source of pleasure and pain, that has a periodic motivational force that is independent of one's capacity for critical judgment" (ibid, p. 76). On his view, all appetites have this feature and at least some kinds of addiction represent disordered, acquired appetites. Part of Watson's aim in comparing the constraints on agency imposed by addiction with those of caring relationships is to argue that we are not justified in disparaging addiction solely because it is a form of dependence (ibid, p. 85). Importantly, his use of "attachment" does not refer to the sense of attachment at issue in this paper.
[21] Frankfurt 1999b, p. 170, Frankfurt's emphasis
[22] Ibid.
[23] Frankfurt 1998, p. 83

others serves to structure and/or reflect one's identity.[24] Given this feature of caring, along with its constitutive role in love, it is easy to conclude that caring necessity has great value.

Notice that caring necessity differs from (what I will call) attachment necessity. Caring about another involves a need for that person to flourish, a need which in turn can give rise to other felt needs to promote the person's good or to prevent her injury. Attachment necessity is not centrally concerned with the *flourishing* of another. When one is attached, what one needs is engagement with a particular person. Attachment necessity involves needing the *other person* in a more direct sense, and as noted earlier, is importantly more self-regarding than caring necessity.

Also, when our attachment needs go unmet, we are subject to a particular type of harm, a reduced sense of security. On most accounts, the central type of harm to which we are subject in caring is emotional pain. Recall that we become sad when an object of our care is doing poorly, fearful when it is in danger, etc. Emotional harm of this sort does not invariably take the form of reduced security. I can feel bad for another for whom I care without feeling as though my own well-being or agential competence is threatened in any significant sense.[25]

Finally, when caring takes the form of love, one's felt need for the cared-for object to flourish – and the volitional necessities arising from this care – become bound up with one's identity in particular ways. This phenomenon has been characterized in terms of integrating the object of one's care into one's identity, identifying one's own interests with those of the cared-for object, or a kind of volitional endorsement of one's attitudes toward the object of care. These particular relations need not obtain between our own identities and (the objects of) our attachment needs.[26]

Thus, even while caring necessity doubtless has value, there might be reason to be skeptical of the value of attachment necessity, as it is more self-regarding, implicates the agent's sense of security (which might be considered a particularly unfavorable form of dependence), and needn't be tied to the agent's identity in the ways that caring is.[27]

[24] For more on the relationship between caring and identity, see Shoemaker (2003) and Jaworska 2007a/b.
[25] This, of course, is not to deny that in some instances, severe emotional distress can become so overwhelming as to dislodge one's sense of security, correctly registering serious threats to one's well-being and agency
[26] Of course, sometimes our attachment objects *are* connected to our identities in these ways, and when this occurs, the relationship is richer and more meaningful in virtue of this connection. Thanks to Ruth Chang for instructive discussion on this topic.
[27] Frankfurt seems to tie the value of love to features of caring, which on my account, attachment does not have. He writes, "I would like to be able to explain just why it is that loving has this intrinsic value. We need to understand

In this section, I reviewed several worries associated with felt necessity, and I explored how these worries might be mitigated in the case of caring necessity, though not necessarily in the case of attachment necessity. With the preceding descriptions and distinctions in hand, I will now offer an argument for the value of attachment.

## §3. The Value of Experiencing Another as a Felt Attachment Need

Attachment, while not synonymous with caring, can be a rich source of value on account of the type of felt necessity that is internal to it. Interestingly, though the value of attachment cannot be inferred directly from that of caring, the two attitudes nonetheless complement one another's respective value in interesting ways.

Let's begin by returning to the nature of vulnerability and dependency that we began discussing in section 2. Research suggests that attachment relationships can provide the raw material for the development of empathy and respect for the value of others.[28] One reason might be that attachment provides us with experiential knowledge of our own vulnerability and dependence, and some theorists have argued that appreciating one's own vulnerability and dependence can, and often does, facilitate moral community with others. Erinn Gilson, for example, argues that experiential knowledge of one's own vulnerability is a requisite starting point for ethical responses to vulnerabilities in others.[29] Recognizing vulnerability and dependence as central features of our own lives allows us to see others' vulnerabilities as evidence of a shared condition between us. Taking up this shared condition is thought to be key to motivating caring attitudes and behaviors toward those who require aid.[30] According to attachment theorists, our capacities for empathy and other forms of moral engagement are heavily influenced by our earliest attachment experiences, and they continue to develop throughout adulthood and may be shaped (partly) by interactions with subsequent attachment figures.[31] In this way, both Tommy's attachment to his mother and Adam's attachment to Linda might help the attached agents become better carers and more capable moral agents.

just what it is about loving that accounts for its importance in our lives. Presumably, the explanation has something to do with the complex fact that loving entails both volitional constraint and disinterested identification with the well-being of the beloved" (1999b, 173-174).

[28] See, for example, Saltaris (2002); Mikulincer et al (2005); Moll (2007), esp. pp. 7-8; Mikulincer & Shaver (2010), pp. 271-273.

[29] Gilson (2014, p. 179)

[30] See Gilson (2014, p. 179); Sarah Clark Miller (2012, p. 8); and Alasdair MacIntyre (1999).

[31] See for example Saltaris (2002) and Mikulincer and Shaver (2007, esp. ch. 2).

Another reason that needing someone can be valuable is that it provides a person with the opportunity to be needed. While we do not *choose* whom we need, and thus do not need them for their sakes, the fact that we need others is often good for them. When an individual feels needed, the experience can lend a sense of purpose to her life. And oftentimes, feeling unneeded can leave an agent feeling as though her relationship is impoverished. Consider the painful lamentations of parents who feel unneeded by their children. Some theorists have associated parents' reluctance to "let go" of children who are approaching adulthood with a continuing desire to be needed by them. As children grow more independent, some parents "panic" and are plagued by such questions as "Who am I if my child doesn't need me anymore?"[32] Interestingly, parents of children who have difficulties forming attachments often report being very disturbed by the fact that their children do not seem to *need* them except in a very superficial sense.[33]

Being needed has import for us because it means that we matter in a very significant respect, and this is especially true in the case of attachment. When someone is attached to you, you alone can fulfill that person's particular need.[34] In this way, you are singularly, or uniquely, valuable to the attached agent. Thus, in being attached to others, we are often providing a source of significant value for them. Being needed, especially by someone whom we desire to benefit, can imbue our lives with a particularly rich kind of value. And when reciprocated, it can deepen and enhance a relationship by fostering intimacy, trust, and a mutual appreciation for one another as uniquely valuable agents.[35] Such relationships are typically good for both parties.

Consider again Adam and Linda. Their mutual attachment marks one another out as uniquely important for how they feel about themselves and how they are able to get on in the world. This orientation, though somewhat self-regarding, affords the relationship a kind of depth

[32] Elisa Morgan and Carol Kuykendall (1997, p. 201)

[33] For example, an anonymous mother of several children with reactive attachment disorder writes, "Even surrounded by people, I feel alone…They don't need me…The kids don't ask for [affection], they don't ask for anything. Ever. Unless it's the newest trinket or toy that they'll die if they don't have" (Annymous 2014).

[34] When you are attached to someone, only *that* person can contribute to your sense of security or well-being in the way that she does. While you might be attached to several persons or objects, each plays a unique role in the type of fulfillment that it provides. In psychologist Mary Ainsworth's words, "…an attachment figure is never wholly interchangeable with or replaceable by another…" (1991, p. 38).

[35] Psychological research on adult attachment suggests that by serving as mutual attachment figures for another, romantic partners can foster trust and closeness in their relationship (Collins & Feeney 2000; Collins et al 2006).

it would not otherwise have and one that does not seem to arise from caring alone.[36] To see this, consider a modified example that I borrow from a previous work.

> Imagine now that Adam is not attached to Linda but that he cares for her selflessly and that in their interactions, he is exclusively concerned with how best to promote her welfare. This morning, Linda learned that she has been offered an opportunity to realize her life's dream of embarking on a six-year space mission. Unfortunately, during this time, she would be unreachable to all earth dwellers. When she relays the news to Adam, he is overjoyed and immediately offers to help her pack. When she expresses concern over being apart from him for so long, he cheerfully replies, "This is what's best for you, and that's really all that matters to me!" [37]

While Adam's attitude in this imagined scenario would certainly be one of caring, we would understand if Linda were disappointed at his response – preferring that he not only need for her to flourish but that he also need *her* (in the sense exemplified by attachment).[38]

The point is not that caring is unimportant, but rather that both attachment and caring do formative work in Adam and Linda's relationship. For example, Linda's care for Adam entails a substantial emotional investment in his good. Adam's need of Linda, in virtue of his attachment to her, means that Linda is not only able to contribute to his good but is, in a significant respect,

[36] It is common for theorists to regard the most valuable types of caring as disinterested. Frankfurt is explicit about this in the case of love, insisting that love be devoid of any self-regarding motives (1999a) Importantly, Frankfurt doubts that romantic relationships provide "especially authentic" paradigms of the kind of caring exemplified by love. He explains, "…the attitudes of romantic lovers toward their beloveds are rarely altogether disinterested, and those aspects of their attitudes which are indeed disinterested are generally obscured by more urgent concerns that are conspicuously or covertly self-regarding" (1999b, p. 166).

[37] Wonderly forthcoming, 11. Note that the point is not that Adam should ask Linda to stay, but rather that in some kinds of relationships, the prospect of our beloveds selflessly devoting themselves to our welfare without ample regard to how engagement with us impacts them and their lives is an unpleasant one. We want them to recognize both cognitively and emotionally the import that we have *for them*, and this would seem to require something beyond caring alone.

[38] To see that we value needing others (and being needed) in this way, consider Dan Moller's remarks on the empirical finding that people are more resilient to being negatively impacted by the deaths of loved ones than typically supposed. According to Moller, "We like to believe that we are *needed* by our husband or wife and that consequently losing us should have a profound and lasting effect on them, just as the sudden injury of a key baseball player should have a disruptive and debilitating effect on the team" (2007, 309). On his view, most of us tend to think that our deaths "would make a deep impact on [our beloveds'] ability to continue to lead happy worthwhile lives," and our beloveds' resilience to losing us would be regrettable because it "shows that we don't have the significance that we thought" (ibid). On my view, this is exactly right. The point is not that we enjoy the thought of our beloveds suffering or being dramatically impaired without us, but rather that the need for us (in the sense associated with attachment necessity) – and not just the need for our flourishing – is often an important element of kinds of caring relationships.

*a part of it*. Being singularly important for him in this way makes Linda's life more meaningful. And because she cares so much for him, she will tend to be especially responsive to Adam's need of her. By regularly and positively attending to this need, Linda can help prevent much of the suffering to which Adam's attachment renders him vulnerable. He is less likely to experience a reduction in felt security due to separation from, or rejection by, her.[39] And the same is true with respect to Linda's attachment to Adam and his care for her.

One might wonder whether cases in which an adult doesn't care about her attachment figure – e.g., Adam and trainer Trent, from section 1, can really have value.[40] I think this is an interesting question, and I'll conclude with a few brief remarks about it. I suspect that relationships like these can have a special kind of significance for both parties and make their joint aims more fruitful. Though Adam's attachment to Trent is restricted to a rather limited domain, there are other examples in which such attachments might have a broader and more significant impact on one's life. Consider that psychotherapists sometimes encourage their patients to become attached to them (without caring in the relevant sense).[41] Such a therapist isn't concerned to cultivate the patient's investment in the therapist's own well-being, but to foster a healthy attachment through which the two can establish a kind of trust and professional intimacy that makes their sessions more productive.

~

In sum, attachment can be a rich source of value because needing others in this way can serve to cultivate a sense of moral community with, and respect for, persons in general, to provide specific others with the opportunity to be needed, and to enhance personal relationships by facilitating a unique and important brand of closeness. Exploring this point has revealed that while distinct attitudes, caring and attachment can complement each other in interesting ways.

[39] There is also evidence that a caring response from one's attachment figure can reduce the negative impact of diminished self-control by increasing one's self-confidence and self-reliance. Supportive interactions with attachment figures imbue us with senses of security, competence, and self-worth that enables us to persevere through difficult circumstances even without the physical presence of our attachment figures (Mikulincer & Shaver 2007). Such interactions, both in infancy and adulthood, also help to shape our internal models of our selves (Bowlby 1969/1980; Mikulincer & Shaver 2007; Karen 1998). In this way, it is conceivable that one's attachment figure can have a more direct and active role in shaping one's agency than another for whom one merely cares.

[40] One obvious way in which attachment (including attachments of this kind) would seem to have value, is that via its third key mark, attachment provides the attached agent with extra resources to cope with adversity and to thrive in ways we might be otherwise unable to. Our attachment figures are uniquely positioned to see us through tough times and to inspire us to take risks and to accept new challenges (Bowlby 1969/1980; Ainsworth 1991; Collins et al 2006).

[41] See, for example, Mikulincer and Shaver (2006, chp. 14).

**References**

Ainsworth, M. 1988. "On Security." From the Proceedings of the State University of New York, Stony Brook Conference on Attachment.

_____. 1991. "Attachment and Other Affectional Bonds Across the Life Cycle," in C.M. Parkes, J. Stevenson-Hinde, and P. Marris (eds.), *Attachment Across the Life Cycle*, New York: Routledge, 33-51.

Annymous. (2014, December 1). My R.A.D. Life. Retrieved from http://rad-narayt.blogspot.com/

Bentham, Jeremy. 2007. *An Introduction to the Principles of Morals and Legislation*. Mineola, N.Y: Dover Publications.

Blatz, W. 1966. *Human Security: Some Reflections*. Toronto: University of Toronto Press.

Bowlby, J. 1969. *Attachment and Loss, Vol. 1: Attachment*. New York, NY, US*:* Basic Books.

_____. 1973. *Attachment and Loss, Vol. 2: Separation*. New York, NY, US*:* Basic Books.

_____. 1980. *Attachment and Loss, Vol. 3: Loss, Sadness, and Depression*. New York, NY, US*:* Basic Books.

Brumbaugh, C., and Fraley, R. 2006. "The Evolution of Attachment in Romantic Relationships," in M. Mikulincer and G.S. Goodman (eds.), *Dynamics of Romantic Love: Attachment, Caregiving, and Sex*, New York: Guilford Press, 71-101.

Burroughs, W. S. 1959. *Naked Lunch.* New York: Grove Weidenfeld.

Christman, John. 2015. "Autonomy in Moral and Political Philosophy." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2015. http://plato.stanford.edu/archives/spr2015/entries/autonomy-moral/.

Cicero: See King 1927.

Collins, N., and Feeney, B. 2000. “A Safe Haven: An Attachment Theory Perspective on Support Seeking and Caregiving in Intimate Relationships,” *Journal of Personality and Social Psychology*, 78(6): 1053-1073.

Collins, N., Guichard, A., Ford, M., and Feeney, B. 2006. “Responding to Need in Intimate Relationships: Normative Processes and Individual Differences,” in M. Mikulincer and G.S. Goodman (eds.), *Dynamics of Romantic Love: Attachment, Caregiving, and Sex*, New York: Guilford Press, 149-189.

Crisp, R. 2014. “Well-Being.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2014. http://plato.stanford.edu/archives/win2014/entries/well-being/.

Dıgha Nikaya. (1987). *The Long Discourses of the Buddha: A translation of the Dıgha Nikaya* (M. Walshe, Trans.). Boston: Wisdom Publications.

Epictetus. (1983). *The Handbook of Epictetus* (N. White, Trans.). Indianapolis: Hackett Publishing.

Frankfurt, H. 1998. "The Importance of What We Care About." In his *The Importance of What We Care About*. New York: Cambridge University Press, 80-94.

_____. 1999a. “Autonomy, Necessity, and Love,” in *Necessity, Volition, and Love*, Cambridge: Cambridge University Press, 129-41.

_____. 1999b. “On Caring,” in *Necessity, Volition, and Love*, Cambridge: Cambridge University Press, 155-80.

_____. 2004. *Reasons of Love*, Princeton: Princeton University Press.

Gilson, E. 2014. *The Ethics of Vulnerability: A Feminist Analysis of Social Life and Practice*. New York: Routledge.

Govier, T. 1993. “Self-Trust, Autonomy, and Self-Esteem.” *Hypatia* 8(1): 92-120.

Graham, A.C. trans. 1981. *Chuang-Tzu: The Inner Chapters*, Indianapolis: Hackett Publishing.

Griffin, J. 1986. *Well-Being: Its Meaning, Measurement, and Moral Importance*. Oxford University Press.

Hazan, C., Campa, M., and Gur-Yaish, N. 2006. “What is Adult Attachment?” in M. Mikulincer and G.S. Goodman (eds.), *Dynamics of Romantic Love: Attachment, Caregiving, and Sex*, New York: Guilford Press, 47-70.

Helm, B. 2010. *Love, Friendship, and the Self: Intimacy, Identification, and the Social Nature of Persons.* Oxford: Oxford University Press.

Jaworska, A. 2007a. "Caring and Full Moral Standing," *Ethics* 117(3): 460-497.

_____. 2007b. "Caring and Internality." *Philosophy and Phenomenological Research* 74(3): 529-568.

Karen, R. 1998. *Becoming Attached: First Relationships and How They Shape Our Capacity to Love*. Reprint edition. New York: Oxford University Press.

King, J. E., (1927/1887). *Tusculan disputations*. Cambridge, MA: Harvard University Press; London: William Heinemann Ltd.; Loeb Classical Library: Latin with facing English translation.

Kraut, R. 2009. *What Is Good and Why: The Ethics of Well-Being*. Cambridge, Mass.: Harvard University Press.

MacIntyre, A. 1999. *Dependent Rational Animals: Why Human Beings Need the Virtues*. Open Court.

Maslow, A.H. 1942. “The Dynamics of Psychological Security-Insecurity,” *Journal of Personality* 10(4): 331-344.

Mikulincer, M. 2006. “Attachment, Caregiving, and Sex within Romantic Relationships: A Behavioral Systems Perspective,” in M. Mikulincer and G.S. Goodman (eds.), *Dynamics of Romantic Love: Attachment, Caregiving, and Sex*, New York: Guilford Press, 23-46.

Mikulincer, M., and Shaver, P. 2007. *Attachment in Adulthood: Structure, Dynamics, and Change*, New York: Guilford Press.

_____. 2010. “Does Gratitude Promote Prosocial Behavior? The Moderating Role of Attachment Security.” In *Prosocial Motives, Emotions, and Behavior: The Better Angels*

*of Our Nature*, edited by M. Mikulincer and P. R. Shaver, 267–83. Washington, DC, US: American Psychological Association.

Miller, S. C. 2012. *The Ethics of Need: Agency, Dignity, and Obligation*. New York: Routledge.

Moll, J., R. de Oliveira-Souza, R. Zahn, and J. Grafman. 2008. "The Cognitive Neuroscience of Moral Emotions," in W. Sinnott-Armstrong, ed. *Moral Psychology: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, vol. 3, Cambridge, MA: The MIT Press, 1-18.

Moller, D. (2007). Love and Death. *Journal of Philosophy* 104(6): 301-316.

Morgan, C., and E. Kuykendall. 1997. *What Every Child Needs: Meet Your Child's Nine Basic Needs for Love*. 1st edition. Bondfire Books.

Rawls, J. 1999. *A Theory of Justice*. Revised edition. Cambridge, Mass: Belknap Press.

Rholes, S., and J. Simpson. 2004. *Adult Attachment: Theory, Research, and Clinical Implications*. New York: Guilford Press.

Saltaris, C. 2002. "Psychopathy in Juvenile Offenders: Can Temperament and Attachment Be Considered as Robust Developmental Precursors?" *Clinical Psychology Review* 22(5): 729-752.

Seidman, J. 2008. "Caring and the Boundary-Driven Structure of Practical Deliberation," *Journal of Ethics and Social Philosophy* 3(1): 1-36.

Shoemaker, D. 2003. "Caring, Identification, and Agency." *Ethics* 114(1): 88-118.

Spitz, R.A. 1945. "Hospitalism: An Inquiry into the genesis of psychiatric conditions in early childhood." *Psychoanalytic Study of the Child* 1: 53-74.

Spitz, R.A and K. M. Wolf . 1946. "Anaclitic depression; an inquiry into the genesis of psychiatric conditions in early childhood, II." *Psychoanalytic Study of the Child* 2: 313-342.

Stampe, D. W. 1988. "Need." *Australasian Journal of Philosophy* 66(2): 129–60.

Van den Berg, A., and K. Oei. 2009. "Attachment and Psychopathy in Forensic Patients." *Mental Health Review Journal* 14(3): 40–51.

Watson, G. 2004. "Disordered Appetites: Addiction, Compulsion, and Dependence." In *Agency and Answerability: Selective Essays*. New York: Oxford University Press, 59-87.

Wiggins, D. 1998. "What is the force of the Claim that One needs Something?" in G. Brock (ed.), *Necessary Goods*, Lanham, MD: Rowan & Littlefield.

Wiggins, D., and S. Dermen. 1987. "Needs, Need, Needing." *Journal of Medical Ethics* 13(2): 62–68.

Wollheim, R. 1974. "Needs, Desires and Moral Turpitude." *Royal Institute of Philosophy Lectures* 8: 162–79.

Wonderly, M. 2016. "On Being Attached." *Philosophical Studies*, 173(1): 223-242.

_____. Forthcoming. "Love and Attachment." *American Philosophical Quarterly*.

Aleksy Tarasenko-Struc
Harvard University

**Bio:** Aleksy Tarasenko-Struc is a PhD candidate at Harvard University. He works mainly in moral philosophy—specifically, on issues connected with moral skepticism, the nature of ethical objectivity, the recognition (and misrecognition) of persons, and the character of personal relationships. He is writing a dissertation which argues that conceiving of moral requirements as grounded in relations between persons provides both a way of formulating a forceful skeptical challenge to morality's authority and a clue to answering it.

# The Kantian Conception of Obligation and the Directedness Constraint

## 1. Introduction

It's a significant fact about morality that at least some of its requirements specify forms of treatment that are *owed to* persons. I will call these *directed obligations*.[1] For example, intuitively we owe it *to others—inter alia—*not to kill them or cause them undue suffering; not to coerce, manipulate, or deceive them; not to interfere with their projects; and to provide them with certain minimal forms of positive aid.

We owe it to others to treat them in certain ways in the sense that we are *accountable* to them for not doing so; in failing to do so, we are not just committing wrongdoing, we are *wronging* them. This means that others have the authority or standing to demand these forms of treatment from us, and perhaps also to blame us if we don't comply. It's even fair to say that some of these obligations correspond to *rights* that we have not to be treated in certain ways. Indeed, the language of common-sense morality appears to be not just 'fraught with ought', but, so to speak, 'rife with right', as well.

If moral requirements include obligations to others, however, this places a significant constraint on what counts as showing that a moral requirement has authority over us. To be successful, any account of a directed moral obligation's validity must show not just that the obligations exists, but also that it really is owed to the one to whom it seems to be owed: *the other per-*

[1] Equivalent terms in the literature are 'bipolar' and 'relational.'

*son*. I will call this requirement the *directedness constraint*. An account would fail to meet it if it yielded the conclusion that our moral obligations to others are, in fact, owed to everyone, to ourselves merely, or to no one at all.

In this paper, I argue that certain contemporary Kantian views cannot meet the directedness constraint. I focus on the work of Christine Korsgaard.[2] The problem I raise for her is that, while she acknowledges the fact that others can obligate us, her conception of obligation makes it puzzling how they could. This is because, on her view, we have an obligation to act in a certain way just when (and because) we have obligated ourselves to act in that way, by adopting the relevant maxim as a universal law. Yet if, on the Kantian story, my obligations ultimately arise from *my* obligating *myself*, it seems that *others* couldn't genuinely obligate me—in which case no moral obligations can be directed after all.

Korsgaard is not insensitive to this sort of concern. Indeed, in later work she tries to fill in her view of the source of our obligations with the thought that parties to interaction reciprocally grant authority to one another.[3] This appeal to *joint self-binding*, as I will call it, seems to allow her view to meet the directedness constraint, as it implies that interlocutors stand in an accountability relation that gives them the authority that their being under directed obligations implies. However, I argue that the appearance is, ultimately, illusory: as formulated, this addition to her account in fact *assumes* a prior accountability relation—the moral relationship—that makes it possible for us to obligate one another. And it's unclear what could explain why, on Korsgaard's view, we must enter into this relation.

Finally, I consider a recent argument of Kyla Ebels-Duggan's which seems well-suited for filling this gap in Korsgaard's account.[4] Ebels-Duggan contends that it is rational or reasonable for us to enter the moral relationship because, insofar as we pursue ends that require the aid or non-interference of others, we must see our choices as giving others reasons for action, on pain of compromising our autonomy. But, given that she construes the relevant conditions of autonomy merely as our having reasons to stick to our projects, her argument is unable to secure such a strong conclusion. I conclude that the conception of obligation held by Kantians cannot meet the directedness constraint. For the Kantian, other persons are not sources of, but merely occasions for, the claims they make on us.

[2] Korsgaard (1996a), pp. 275-310; (1996b), pp. 132-145; and (2009), pp. 177-206.
[3] Korsgaard (2009), pp. 177-206.
[4] Ebels-Duggan (2009), pp. 1-19.

**2. Korsgaard's Conception of Obligation and the Directedness Constraint**

Korsgaard presents her account of the reality of others' claims in Lecture 4 of *The Sources of Normativity*.[5] There she advances considerations to show that reasons for action are *public* rather than *private*, in her terms. Roughly, a reason is public just when it is such as can be given by one person to another in the space of interaction, so that the reason comes to be shared by both parties—and private otherwise. And, she argues, reasons are by their very nature public. She clarifies what, in her view, this sharing of reasons consists of in her brief remarks concerning the normative force of interpersonal address:

> If I call out your name, I make you stop in your tracks… Now you cannot proceed as you did before. Oh, you can proceed, all right, but not just as you did before. For now if you walk on, you will be ignoring me and slighting me. It will probably be difficult for you, and you will have to muster a certain active resistance, a sense of rebellion. But why should you have to rebel against me? It is because I am a law to you. By calling out your name, I have obligated you. I have given you a reason to stop.
>
> Of course that's overstated: you don't have to stop. You have reasons of your own… But that I have given you a reason is clear from the fact that, in ordinary circumstances, you will feel like giving me one back.[6]

In addressing another person, we give her a reason to respond to us in the way we are calling for, one that cannot simply be ignored without personally affronting us. Of course, it may in fact be just *a* reason: she may have a reason of her own to decline to respond in this way, which she may in turn express to us. The aim of the interaction is to arrive at reasons that both parties can basically accept. This feature of address is manifest in cases of coordination or collective deliberation, in which the responses of our addressee are normally taken as directly generating reasons for us as well. That, for example, the person with whom we are trying to schedule a meeting can't make a certain time is automatically a reason for us to find a different time, too, one that will be acceptable to both parties.[7]

[5] Korsgaard (1996b), pp. 132-145.
[6] Ibid., p. 140.
[7] Ibid., p. 141-142. See also Korsgaard (2009), pp. 192-196.

However, Korsgaard also suggests that it is not only address that manifestly involves reason-sharing of this sort but interaction more generally as well:

> We do not seem to need a reason to take the reasons of others into account. We seem to need a reason not to. Certainly we do things because others want us to, ask us to, tell us to, all the time. We give each other the time and directions, open doors and step aside, warn each other of imminent perils large and small. We respond with the alacrity of obedient soldiers to telephones and doorbells and cries for help.[8]

In some of these cases (opening doors, stepping aside, warning one another of imminent perils), others arguably make claims on us, but address proper is not involved. Instead, we are responding directly to the unaddressed reasons of others, thereby treating them as public, as reasons for us. Korsgaard's point is that this form of responsiveness to the reasons of others is our default stance. Indeed, according to her, not only is it part of what defines our social nature as human beings, it is also what makes it possible for us to share a social world in the first place. And in sharing a social world with one another, we are committed to recognizing that our reasons are shareable in this sense as well.

Taken together, Korsgaard's remarks suggest a phenomenologically accurate view of making a claim on another person (or obligating her), one that highlights two of its significant features. First, when one person makes a claim on another the claim is made *directly*, in the sense that the particular other herself is the source of, rather than merely the occasion for, the claim's being generated. When you succeed in obligating me, the ultimate ground of that obligation is *your* will—which is to say, *you*—while *my* will does not appear to be so intimately involved in the obligation's generation. And this is a second, related feature of interpersonal claim-making: it is, or seems, *non-voluntaristic*. In general, it does not seem that claims are valid only through an act of my will. I do not, as it were, need to voluntarily *validate* a claim for it to be valid; rather, I respond to it, in an unmediated way, *as valid*.

Yet this picture is difficult to integrate with Korsgaard's general conception of obligation, which suggests that her view cannot meet the directedness constraint.[9] According to her, the

[8] Korsgaard (1996b), p. 141.

[9] Ibid., pp. 92-93. See Korsgaard (2007), pp. 10-11. On her conception of self-binding, it also involves not just authority over ourselves but accountability to ourselves as well. On Korsgaard's view, we are accountable to ourselves in the sense that we can legitimately demand an explanation from ourselves of what we are doing or what we have chosen.

origin of our obligations lies in self-binding: when we incur obligations, it is when and because we have obligated ourselves to act accordingly. And this happens when we choose to act for the sake of an end, she thinks, thereby adopting the relevant maxim—of acting on that end—as a universal law. Hence, the source of obligation—and thus of reasons for action—is our own authority over ourselves, the authority to give ourselves binding laws, which we exercise just by choosing ends at all.

But now the claims of others seem to be valid simply virtue of these others, their activity or their plight. Another way of saying this is that others are *themselves* the source of, not merely the occasion for, our obligations to them; in particular, the ground of their validity does not appear to lie in any act of my will. Yet if Korsgaard's general view of the source of obligation is true, an act of my will is surely what makes the claims of others valid: they are validated just when I give myself the law of acting on them. I am not obligated to act unless I obligate myself. But, in that case, how can others obligate me, and indeed, obligate me in the sense that they constitute the *source* of my obligation?

The upshot for the Kantian picture seems to be that all moral obligations to others are not actually owed to them at all but are in fact owed to *oneself*. An example will help to illustrate the nature of the challenge. Imagine a view—let's call it *naïve voluntarism*—on which all of an agent's obligations derive from promises she has made to herself. The only basic obligations are promissory obligations to the self. This view seems to have the strange consequence that we never really owe it to *others* to treat them in certain ways; instead, we always merely owe it to *ourselves* to treat them in those ways. For example, suppose you are drowning in a lake and, seeing you, I promise myself that I will rescue you. If I fail to keep this promise, according to the naïve voluntarist, I'm entitled to demand an explanation from myself and reproach myself, but the other has no such standing to do these things: I'm not accountable to the other, I'm accountable only to myself since it's *I* who have obligated myself. You were only the occasion for my obligation, not its source, and so the obligation was not *to you*. The crucial question here, then, is this: Why doesn't Korsgaard's view incur the same false conclusion? If grounding the validity of others' claims in acts of self-promising rules out our having obligations to others, why shouldn't grounding it in acts of self-binding have exactly the same effect?

So, if all of an agent's obligations derive from her acts of self-binding, as on Korsgaard's picture, the result is that all the moral obligations that we thought were owed to others are all ul-

timately only owed to the self, not the other. Her Kantianism cannot meet the directedness constraint, it seems.

**3. Joint Self-Binding as the Basis for an Escape Route**

Korsgaard has an obvious and powerful line of response available to her. Kantianism, she might say, only seems not to meet the directedness constraint because we've assumed that for the Kantian the subject of self-binding is the *single individual*: that all of an agent's obligations are ultimately grounded in her acts of obligating *herself*. But the view need not be committed to holding this. Certainly, Kantianism is best interpreted as affirming that the subject of self-binding is *rational agents generally* and so that at least some of an agent's obligations derive not from her binding herself to act in some way, but from her and others collectively binding themselves to do so, which I will call *joint self-binding*. It might be thought that appeal to this idea enables the Kantian to meet the directedness constraint.

Let's first get Korsgaard's conception of joint self-binding into view.[10] Joint self-binding is the activity of making laws together, and her conception of it is explicitly modeled on Kant's view of personal relationships such as friendship and marriage. In standing in these relations, she claims, we make a kind of *joint commitment*.[11] That is, you and I reciprocally cede our unilateral authority with respect to some range of choices that affect either, or both, of us. I cede my authority to make these choices on my own *to you*, and you in turn cede your authority to make similar choices on your own *to me*. In so doing, we commit ourselves to arriving at shared decisions on these matters, decisions we arrive at by practical deliberation that we engage in together. The aim of this joint deliberation is the free choice of a law valid for both parties, the construction and realization of a common good.

How could this idea help Korsgaard meet the directedness constraint? One answer is suggested by her remark that promises and agreements consist in an act of joint self-binding, so construed, as does interaction more generally.[12] Like friendship and marriage, she thinks, entering into the promisor-promisee involves the reciprocal ceding of authority. If I promise you that I'll pick you up from the airport at a certain time, I thereby give up my authority to choose whether or not I come to the airport then, and I grant it to you, making myself accountable to you for

[10] The discussion is scattered across Korsgaard (2009), pp. 186-202.
[11] Or, as Korsgaard puts it, following Kant, 'a unity of will' or 'the formation of unified wills.' Ibid., p. 187, 190.
[12] Ibid., pp. 189-190.

compliance. You are then entitled to be picked up from the airport by me at that time, and you may legitimately ask me to account for my behavior if I don't. Similarly, you give up your authority to choose whether or not to remain at the airport at the designated time; I could demand an explanation of your behavior if, say, you left the airport before I came to pick you up. Now if on Korsgaard's account all interaction has this structure, maybe the proposed solution would be that, when we interact with one another with a view to making a claim, we thereby make a joint commitment to deliberating on matters that affect either, or both, of us—treating these matters as calling for a shared decision acceptable to all parties.

If all *claim-addressing interactions*, as I will call them, are analogous in these respects to the relation between promisor and promisee, the problem seems to dissolve. Whenever we make claims on each other in interaction, we grant one another the authority we normally enjoy over our own choices and make ourselves accountable to one another for how we then act. There is therefore no puzzle, on this line, why some of our moral obligations are owed to others instead of to ourselves. These obligations are owed to others because we have granted others the authority to obligate us in certain ways.

There are two problems with the proposal that claim-addressing interactions are relevantly similar to the promisor-promisee relation, however. One is that making a promise is a voluntary act, which we may or may not perform, while in many cases recognizing another's claim is not. Promissory obligations to others are only generated when the promise is actually made to her: the existence of the obligation depends on an act of voluntary ratification on the part of the would-be promisor. This means that, if she does not make a promise, she is not obligated to act in the relevant way. Making a promise, then, is like giving someone a gift in person: for a gift to be given by one party, it must be taken by another party. But claim-making interactions do not seem to be this way. You can make a claim on me without my assent, so to speak. For example, if I walk by and see you drowning, you make a valid claim on me even if I don't first 'accept' your condition as reason-giving and validate it—as if in your distress you were offering me a contract to sign and stamp. This is a significant disanalogy.

Yet there's also a second, deeper problem. Promises, like claim-making interactions in general, actually *presuppose* a background authority relation between the parties, which determines whether the claim so made is valid or invalid. Let's return to our earlier example. If you call me and ask me to pick you up from the airport tomorrow, the validity of your claim depends

on the character of the relationship between us. If we are friends, you will have a claim on me that I pick you up from the airport then, but you will not if, say, we are perfect strangers. So, even prior to and apart from the interaction, there must be some relation of authority and accountability between us if your claim on me is to be valid. If certain claim-making interactions generate directed moral obligations for us, then if morality binds everyone in virtue of the fact that we are persons, there is an important consequence: for some of the claims we make on one another to be valid, there must be a more generalized normative relation between us, in virtue of which these claims are valid. We must therefore stand in what we might call the *moral relationship* to one another—a relationship of authority and accountability between every person and every other. Standing in this relation gives us each the power to obligate one another in certain ways—that is, to call for certain forms of treatment from one another with recognizable legitimacy.

But then Korsgaard has not shown why we should enter the moral relationship in the first place, nor has she shown that we need not answer that question (perhaps because it is moot or incoherent); she has, in effect, presupposed that there is such a relation. And she does need a story about why we stand in the moral relationship with others to begin with, lest her account end at an arbitrary stopping point. Also, it's unclear what resources the Kantian view has in this regard, given its emphasis on the significance of first-personal willing: its view that, as Kant says, 'I can recognize that I am under obligation to others only insofar as I at the same time put myself under obligation.'[13] At any rate, there is now no explanation, in Korsgaard's account, of why we must engage in claim-making interactions to begin with; this seems now like an open and indeed legitimate question, which goes unanswered. Korsgaard's appeal to joint self-binding cannot by itself enable her to meet the directedness constraint.

**4. Ebels-Duggan's Argument for the Rationality of Entering into Moral Relations**

In a recent paper, Kyla Ebels-Duggan presents a Kantian vindication of the rationality of entering into moral relations, which is supposed to explain how we can have genuine obligations to others even though, on the Kantian view, the source of one's obligations is one's binding oneself.[14] In so doing, she is, in effect, providing one promising way of shoring up the lacuna in

[13] Kant, MM 6:417.
[14] Ebels-Duggan, op cit.

Korsgaard's account. If the issue with Korsgaard's account is that it presupposed an unjustified, unexplained reciprocal authority relation, then perhaps the solution is to show that we do have reason to enter into this sort of relation.

Ebels-Duggan's strategy is to advance an argument modeled on Kant's justification for exiting the juridical state of nature and instituting the civil state. Just as recognizing the authority of a legitimate government capable of adjudicating and enforcing property claims is necessary for our remaining free from interference by others, so too, on Ebels-Duggan's reconstruction, is recognizing that we have the authority to give one another reasons through our choices crucial for our autonomy. This is why we must 'leave' what she calls the *ethical state of nature* and enter into moral relations with others.

But what is the ethical state of nature, exactly, and why is it supposed to be a problematic condition? According to Ebels-Duggan, the ethical state of nature is a condition in which we ascribe no authority to others, not regarding them as making valid claims on us through their choices; the only authority that we recognize is our own authority to make binding laws. The problem is that we *do* still make claims on others. These claims just aren't valid: they assume an authority that we lack. So, we are stuck unilaterally, which is to say invalidly, making claims on one another: we cannot recognize our claims on one another (by acting on them) without compromising our autonomy.[15] But since we are finite, we need to make claims on one another to accomplish many of our ends: we need others not to interfere with our pursuit of our ends, and to provide minimal forms of support.

Ebels-Duggan will argue that inhabiting the ethical state of nature is incompatible with our retaining our inner freedom or autonomy, but first she specifies a condition on its full exercise. Following Kant, she claims that our autonomy is threatened mainly by our inclinations. The threat is especially acute when it comes to temporally extended actions—projects—which are vulnerable to disruption by our yielding to our inclinations to reconsider or drop what we are pursuing. We need reasons to stick to our chosen courses of action in the face of this persistent threat. And for Ebels-Duggan, these reasons have two distinct grounds. The first is the value of our chosen project. We need reasons of this kind to withstand the temptation to pursue other, worthless ends for which we may be inclined. The second ground we need is our choice of this

[15] According to Ebels-Duggan, another problem with the ethical state of nature is that it's unclear just how far the general duty of beneficence extends—how much of a claim others have on my aid, and vice versa—and how one can owe it to a particular person to help her. See Ibid., pp. 9-10, p. 14.

project over others, which makes it possible for us to override inclinations to reconsider our pursuit over others that may be equally valuable. To exercise our inner freedom, then, we must take both grounds to give us reasons to act.

But now Ebels-Duggan more controversially claims that we do, and need to, take these grounds to give *others* reasons to act. First, we must take the *value* of our projects to give others reasons: it is part of the concept of a reason to act that, if you judge that I have a reason, you are committed to acknowledging that anyone similarly situated would have a reason, too. By itself, she says, this is insufficient for autonomy. For the value of our projects cannot be the basis for the common thought that others *owe it* to us not to interfere with these projects, or, relatedly, that their interference is an occasion for personal affront, justifying reactions such as resentment and blame. These attitudes only make sense if we see our choices as giving others reasons independent of the value of what we pursue.

And when we see our choices in this way, Ebels-Duggan maintains, we will in fact make claims on others, calling for their non-interference and a modicum of helpfulness. We *need not* assert either any claims, she admits, but not doing so commits us to an unacceptable restriction: we could only pursue those ends that 'do not require anything, including non-interference, from others'; however, 'given the extent of our interdependence, this restriction is severe,' she says.[16] What follows from this is what she calls the Postulate of Reason Creation: that it must be possible for us to create reasons for one another by setting ends.[17] And, according to her, it's possible through our exiting the ethical state of nature and enter into moral relations, in which we each acknowledge a common authority to give one another reasons through the discretionary choices that we make.

If Ebels-Duggan's argument is successful, then she has given Korsgaard a way of disarming the challenge I have raised in this paper. Is she successful? In fact, I believe that her case is inconclusive. She has not shown that unless I regard my chosen ends as giving others a reason to act, I will not be able to see myself as having sufficient reason to continue pursuing any project I've undertaken. For one, it's unclear why, in order to see myself as having sufficient reason to continue, I need to also see the value of my chosen end as giving others reasons. In support of this point Ebels-Duggan insists that, as a matter of conceptual fact, my acknowledging that you

[16] Ibid., p. 13.
[17] Ibid.

have a reason commits me to judging that I would have a similar reason in your place. This is certainly true, but it does not support her stronger point. Judging that I would have the same reason in your place is apparently compatible with denying that your reason gives me a reason to act—with failing to accord you authority. This is apparent in contexts of competition, in which one party recognize that the other has a reason to seek a certain benefit and, on that basis, concludes that she herself should prevent the other from getting it.

And it's equally unclear why, to be able to regard myself as having sufficient reason to continue my chosen project, I need to see my choice of the project as giving others reasons as well. Why can't I see my choice as giving only myself a reason? This does not seem impossible if all that is required is that I have reasons *not to reconsider* my project. After all, it seems that I can very well stick with my project in the face of inclination with the thought 'I chose to do this'. Ebels-Duggan might now respond that we must see our choices as giving others reasons because we must be able to make claims on their non-interference and their aid; and if we could not do so, then we couldn't stick with our projects in the face of our inclinations. But this too seems unsupported. It seems that I can well engage in apparently claim-making behavior—pleading, imploring, demanding—even if I do not believe that my addressee has the relevant reason. At this point it's unclear why you and I should enter moral relations when we can settle for a pact of mutual non-interference and minimal aid; for the purpose of safeguarding our ability to withstand inclination, that sort of arrangement would work just as well.

I conclude, then, that contemporary Kantianism is left with an unmet challenge: to explain how the first-person practical perspective could possibly ground the authority of the other person.

## References

Ebels-Duggan, Kyla. 'Moral Community: Escaping the Ethical State of Nature'. *Philosophers' Imprint* vol. 9:8, August 2009.

Kant, Immanuel. *The Metaphysics of Morals*, tr. Mary Gregor. Cambridge: Cambridge University Press, 1996.

Korsgaard, Christine. 'The Reasons We Can Share: An Attack on the Distinction Between Agent-Relative and Agent-Neutral Values'. In *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press, 1996a.

——— *The Sources of Normativity*. Cambridge: Cambridge University Press, 1996b.

——— 'Autonomy and the Second Person Within: A Commentary on Stephen Darwall's *The Second-Person Standpoint*'. *Ethics* 118 (October 2007).

———*Self- Constitution*. Oxford: Oxford University Press, 2009.

Berislav Marušić
Brandeis University

**Bio:** Berislav Marušić is Associate Professor of Philosophy at Brandeis University. His research interests lie at the intersection of ethics, epistemology and philosophy of mind. He is the author of *Evidence and Agency: Norms of Belief for Promising and Resolving* (Oxford University Press, 2015), as well as articles on skepticism, action theory, reasons and promising.

# Do Reasons Expire?

My mother passed away on November 30, 2007—suddenly and unexpectedly at the age of 55. In light of this loss, I immediately experienced intense grief. And it seems that my reason for grief was, precisely, this loss—that my mother had died, not young, but too young. Indeed, if I had not experienced such grief, something would have been wrong with me. Contrast me with Camus' character Meursault in *The Stranger* who, a day after his mother's funeral, goes to the movies with a new love interest (1942/2012).

Yet now, many years later, I do not experience much grief, and when I do, it is on particular occasions for particular reasons. Indeed, this, it seems, is as it should be. In "Mourning and Melancholia," Freud puts it with apparent simplicity:

> [A]lthough mourning involves grave departures from the normal attitude to life, it never occurs to us to regard it as a pathological condition [like melancholia, i.e. depression] and to refer it to medical treatment. We rely on its being overcome after a certain lapse of time. (1917/1999, 243-244)[1]

In a similar vein, DSM-5also distinguishes grief from a pathological condition like depression and states that grief normally diminishes with time:

> In distinguishing grief from a major depressive episode (MDE), it is useful to consider that in grief the predominant affect is feelings of emptiness and loss, while in [a major depressive episode] it is persistent depressed mood and the inability to anticipate happiness or pleasure. The dysphoria in grief is likely to

[1] I take "mourning" and "grief" to refer to the same sentiment—what in German is rendered as "Trauer."

> decrease in intensity over days to weeks and occurs in waves, the so-called pangs of grief.[2]

Yet things are not so simple. My grief has passed. But my loss has not been undone. It is still true that my mother died quite young on November 30, 2007. It is still true that this is a loss that I have suffered; the passage of time has not changed that. Indeed, it seems to me that my loss has not diminished in any way. Yet if my grief was a rational response to my loss, and if my loss remains the same over time, then, it seems, I am irrational today—or, better, I am failing to be responsive to my reasons.

However, surely I am not irrational. It is, after all, to be expected that grief diminishes as time passes, because grieving consists in coming to terms with a loss. We might say with Freud, "In mourning time is needed for the command of reality-testing to be carried out in detail, and… when this work has been accomplished the ego will have succeeded in freeing its libido from the lost object" (252). Thus we might say that grief diminishes, because grief consists in work—the work of coming to terms with our loss. And as we complete our work, we have less reason to grieve, or perhaps no reason at all.

The trouble with this suggestion, however reasonable and true to our experience it may be, is that it is hard to see how it could be true. Upon reflection, it seems paradoxical: We want to maintain that our grief is a response to our reasons, that the reasons consist in our loss, and that our loss does not diminish over time, but that it is rational to grieve less as time passes. Yet how could the diminution of grief be rational, if the reasons for grief stay the same? Or how are we to make sense of the thought that reasons for grief expire? Do reasons have an expiration date?

This paradox is not specific to grief. Similar examples can be formulated for other emotions—such as anger and resentment, indignation and shame, regret and affirmation, gratitude, delight and pride. I think it raises especially interesting issues with regard to emotional our response to injustice. However, in what follows, I will focus on grief, and I

[2] American Psychiatric Association (2013). It may be worth mentioning that DSM-5 does not treat ordinary grief as a form of depression, despite its controversial elimination of the "bereavement exclusion" that was part of DSM-4. The point of the bereavement exclusion was to avoid diagnosis of major depression in those who were freshly grieving (with some exceptions). DSM-5 recognizes that depression may be concurrent with grief—indeed that it may be triggered or aggravated by it. But it is a mistake, made frequently in public discourse, that DSM-5 identifies grief and depression or medicalizes grief. Indeed, DSM-5 explicitly states that "Uncomplicated Bereavement" is not a mental disorder.

will other moral sentiments for another occasion—though I do think that my arguments can be extended to them.

As a preliminary, I should note that it is a fundamental assumption of the present approach that grief is, in principle, responsive to reasons. I think that this assumption is warranted, though not unproblematic. It is warranted, because our emotions are not just conditions that befall us, but they partly constitute our take on the world: In fear, we take something to be dangerous, in anger we take something to be a wrong, and in grief, we take something to be a loss.[3]

However, I also want to grant that there is something problematic about the claim that emotions are reasons-responsive. Richard Wollheim puts it well:

> Perhaps, if we are to think of some emotion of ours as altogether rational, we must think of its object as deserving it. But that is neither the norm that our emotions follow, nor one to which we think they should comply. In our emotional life, we do not always feel ourselves to have right on our side. (1999, 115)

It is, indeed, true that in our emotional life, we do not always feel ourselves to have right on our side: We might experience sadness and think that we have no reason to be sad, or we might experience fear and know that there is nothing to be feared. But that we do not always feel right does not suggest that we do not aspire, in our emotional lives, to be responsive to reasons.[4]

In what follows, I will first consider the view that reasons don't expire and argue that it is unacceptable. I will then consider the view that reasons do expire and argue that it, too, is unacceptable. Finally, I will explain why we are faced with a paradox which reveals a principled limit to our understanding of the temporality of our emotions.

## 1. Reasons Don't Expire: The Permanence of Loss

[3] See Solomon (1976), de Sousa (1987), Greenspan (1988), Roberts (1988), and Nussbaum (2001) for defenses of this view.

[4] Wollheim (1999) argues that our emotions provide us with an attitude to the world, in contrast to beliefs, which give us a picture of the world. But it seems to me that an attitude is reasons-responsive: the question why to have an attitude clearly finds application. Thus I think that Wollheim's picture could allow that emotions are reasons-responsive, though perhaps not in the same way as beliefs.

A first response to the paradox I raised insists that reasons don't expire. The main rationale for this view is the plausible observation that a loss does not cease to be a loss as it recedes into the past. The death of my mother was a loss when it occurred and remains a loss to this day, even as I move on in life. It is not undone by the passage of time, and it is not undone by the many events in my life that have occurred since then, such as the birth of my children. But since this loss is a reason for my grief, and since it remains a loss, it remains a reason for grief.[5]

Indeed, we can think of this hardline view as the temporal counterpart to Peter Singer's view in "Famine, Affluence, and Morality" (1972) that spatial distance does not affect our reasons to aid those in need. Presumably, Singer's view applies to time as well as to space. Our temporal distance to others may limit the ways we can aid them, especially if they lived in the past, but our temporal distance as such does not seem to affect the general moral facts, and therefore the appropriateness or rationality of our moral response to them. Just as whether something is a loss or a harm does not depend on *where* it occurs, it does not depend on *when* it occurs. To echo Simone Weil: What's time got to do with it? —"Comme si le temps faisait quelque chose à l'affaire"?[6]

I think that there is something attractive about the hardline view: It is neat and clear and uncompromising. But I do not think that it can be right. Temporal distance does not merely make us grieve less; in many cases it seems to make it *appropriate* to do so.

[5] This may seem especially plausible if there is a close connection between values—such as losses—and reasons: For instance, on a view like T.M. Scanlon's, for something to be a value is for it to have a particular second-order property—namely the property of having properties in virtue of which we have reasons for certain attitudes and actions (1998, 97). And it seems plausible to hold that something's being a loss consists in having a property that provides us with reasons—such as a reason to grieve. On Scanlon's view, we can use the notion of reasons to explain values. But for present purposes—when we are trying to settle what reasons we have—we could see the explanation as going the other way around, from values to reasons.

[6] Maurice Schumann recalls Simone Weil saying to him: "How can we condemn the holocausts which are in preparation or are being perpetrated around us if we don't condemn, or even acknowledge the holocausts as truths of the faith [i.e. the killings described in the Hebrew Bible] under the pretext that they occurred thousands of years ago, *as if time made a difference to the matter*?" (Kahn, ed. 1978, 25, translation and italics mine). ("Comment pouvons-nous condamner les holocausts qui se préparent ou qui se perpètrent autour de nous si nous ne condamnons pas, ou même si nous reconnaissons comme vérités de la foi les holocauses sous prétexte qu'ils se sont écoulés il y a un certain nombre de millénaires, comme si le temps faisait quelque chose à l'affaire?"). I owe the reference to Yourgrau (2010, 127). Weil is reported to have claimed this in 1942, without knowing that the Holocaust was happening and before the term "holocaust" was in use. See Yourgrau (2010) for an account of Weil's own suffering over temporally distant harms.

We know it to be a fact that reasons expire, because we know that it is rational to grieve less after a period of time has passed.

A proponent of the hardline view might respond that our thinking here is subtly confused: She might argue that we confuse the advantageousness of grieving less over time with its rationality. It might be true that we are better off if we are resilient and get over our losses in little time. Indeed, it might be that this is of utmost importance for creatures like us whose lives are fraught with losses. However, it would constitute a reason of the wrong kind to grieve less. It would merely mean that we have good reasons to get ourselves to stop grieving—not reasons which show grief to be unreasonable. Similarly, we might have good reasons to get ourselves to believe or disbelieve something, but those would be the wrong kind of reasons to believe or disbelieve—and so no reasons at all.[7]

Yet I do not think that the objection to the hardline view has to commit this error. It is not just that we are better off if we stop grieving after a short period of time. Rather, it seems that this is a feature of the reasons for grief themselves. There is something wrong with *dwelling* on a loss too long.[8] There is such a thing as being *stuck*. The common psychological term is "complicated grief," and DSM-5 classifies it as "Persistent Complex Bereavement Disorder". Persistent grief seems to somehow be the wrong response to loss; it seems to be irrational. At the very least, persistent grief is not the rational response to loss. The challenge is to explain why that is the case.

**2. Reasons Expire: Circumstances Change**

Let me, then, venture such an explanation. If reasons expire—if it is rational to grieve less as time passes—then this must mean that, as time passes, we gradually have less reason to grieve; and perhaps, in time, we have no reason to grieve at all. This, I

[7] See D'Arms and Jacobson (2000, 77) for a point along these lines about grief and Hieronymi (2005; 2013) for an account of the wrong kind of reasons that I find convincing.

[8] David Owens makes a similar point in the context of anger: "Suppose you commit some significant but not heinous offence against me, and are without excuse. After ten years it still rankles. Perhaps I don't display my continued annoyance but I still feel it. Most would agree that this anger is inapt. I shouldn't bear grudges like this; I should get over it. Even if my grudge is on the whole desirable (as the only thing that keeps me going) it remains inapt. Forgiveness might remain inappropriate (perhaps because it has never been requested) but one can get over a past offence without actually forgiving it and, in many cases, the sheer passage of time renders this appropriate" (2012, 33).

think, is a somewhat cold view. But there certainly is something plausible about it: In time, our circumstances change, and the loss matters less to us. This is clearest if we consider a small loss. For example, suppose I drip ketchup on my favorite shirt and ruin it for good. After a few days or weeks, I settle on a new favorite shirt. At that point, my old shirt doesn't really matter to me: I am over it.[9]

Perhaps things are similar when it comes to our dead loved ones—though of course it takes more time. And even if the dead cannot be replaced, because people are irreplaceable,[10] it is true that, as time passes, the dead loved one plays a less important role in our life. Martha Nussbaum offers an eloquent account of how our grief changes as time passes—though I do not think that she would endorse the cold view.

> When I receive the knowledge of my mother's death, the wrenching character of that knowledge comes in part from the fact that it violently tears the fabric of hope, planning, and expectation that I have built up around her all my life. But when the knowledge of her death has been with me for a long time, I reorganize my other beliefs about the present and future to accord with it. …
> I will still accept many of the same judgments—including judgments about my mother's death, about her worth and importance, about the badness of what happened to her. But propositions having to do with the central role of my mother in my own conception of flourishing will shift into the past tense. … Some things stay constant: my judgments about her intrinsic worth, and about the badness of what happened to her, my judgment that she has figured centrally in my history.… But I put her into a different place in my life, one that is compatible with her being dead, and so not an ongoing active partner in conversation, love, and support. (2001, 80-82)

Nussbaum vividly illustrates how, as time passes, the role that the dead loved one plays in our life changes—and this change may explain why reasons expire.

However, as it stands, Nussbaum's observations do not explain why grief why reasons for grief expire or even significantly diminish at all. That is because, as she puts

[9] Thanks to [ … ] for the example.
[10] See Preston-Roedder and Preston-Roedder's (forthcoming) response to Moller (2007) for thoughtful discussion of this point.

it, some very important things "stay constant"—that one has suffered a loss, that the loss is significant, that one loves the person who died, and that that person matters. Indeed, it seems to me that Nussbaum provides arguments for why grief changes rather than for why it diminishes—why, as she puts it, it becomes a "background emotion" (80). That is why I think that she does not endorse the cold view. And that is also why I think that she does not recognize the full force of the problem of diminishing grief.

The problem is that normal grief does not just change. It diminishes very significantly—to the point that we don't experience it at all, the vast majority of time. To see this, contrast grief with romantic love. Love is first felt very intensely. However, eventually it turns into a less intensely felt, but deeper, emotion that informs much of our thought and action—what Nussbaum might call a background emotion. Indeed, we do not have an expectation that love should expire. We don't think that if you've been with your lover for two months or two years, you should get over it.[11] In the good case, love is permanent. In contrast, grief diminishes and, in the good case, we get over it pretty much completely in fairly little time. This is why we face a paradox: We judge that the death of the loved one matters to us, and matters very much—but we don't respond to it with what would seem to be the appropriate level of grief. Our emotional response does not correspond to our judgment of value.[12] How could this be rational?

Here is a somewhat plausible way to explain why it might be rational to grieve less even though our loss stays constant and continues to matter: It is an important feature of our reasons that they don't have weight or significance in isolation but that their weight or significance depends on other reasons we have in particular circumstances or in a situation.[13] For example, the fact that I have theater tickets to see a show tonight—a show I very much want to see—is an excellent reason to see the show. But if my daughter is very sick, then this reason is attenuated or defeated: Instead I have reason to stay with

[11] The permanence of love is a topic I plan to pursue on another occasion.
[12] Moller makes the point forcefully: "Part of what being the vulnerable creatures of flesh and blood that we are means is that we are subject to staggering losses in the form of the deaths of those we love, and yet our reaction to those losses is utterly incommensurate with their value, especially after the first month or two have passed" (2007, 310).
[13] Scanlon (2014, 30-31) proposes to understand reasons as a four-place relation that holds between a fact, an agent, a set of conditions, and an action or attitude. I take my discussion of circumstances to be equivalent to Scanlon's argument place for conditions.

her and, if things are very bad, to take her to the emergency room.[14] The weight or significance of the fact that I have theater tickets to a show thus depends on my other reasons. And it may be that as our circumstances change over time, our reasons for grief diminish—not because a change occurs in them, but because a change occurs in our other reasons. At the moment of my mother's death, her death matters a great deal to me. But now, many years later, my circumstances are different: I have a family of my own with two little children who place considerable demands on me, and there are many other things I care about which make grieving less urgent. In light of this, it makes sense, and indeed it is rational, that my grief would take up less space in my life—not because my loss has ceased to be a loss, but because my life is filled with other values.

However, even this proposal, as it stands, has a flaw. It makes it too contingent a matter whether reasons expire. That is because what reasons we have, and which circumstances we find ourselves in, is a contingent matter of fact. To illustrate, in somewhat oversimplified terms: The proposal, as it stands, implies that if I had had fewer kids, or fewer important commitments, I would now have weightier reasons to grieve. And if I had had twins the night after my mother died, I would hardly have needed to grieve at all. Indeed, the proposal, as stated, does not attribute any particular significance to the temporal dimension of grief—except insofar, as a contingent matter of fact, it takes time for circumstances to change. But the hypothesis that reasons expire is plausible, because it seems to be in the nature of grief to diminish *as time passes*.

It is possible to address this flaw. To do so, we have to distinguish changes in circumstances that are internal to grief and changes that are external.[15] Changes in circumstances that are external to grief don't have anything specifically to do with grief: How many kids I have is independent of how much I have grieved; my having kids is not my way of grieving. That is why the fact that I have two kids cannot account for my having less weighty reasons to grieve. If I hadn't grieved at all—say, because I learned of my mother's death only now—I would have as much reason to grieve now as if I had the night my mother died, despite the fact that now, unlike then, I have two kids who make considerable demands on me. Hence not just any change in circumstances could explain

[14] See Dancy (2004, ch.2-3) and Schroeder (2007, ch.7) and Horty (2012).
[15] I am grateful to […] for this suggestion and the terminology.

why reasons for grief expire. It has to be a change in circumstances that is internal to grief.

It is difficult to explain what exactly makes a change in circumstances internal to grief. It is not the fact that something is a causal consequence of grief: I might become distracted by grief, break my hip, and end up in the hospital. My circumstances would have changed as a consequence of my grief—but the change would not make it the case that I have less weighty reasons to grieve, despite the fact that I now have my health to worry about. Rather, to capture the point that the change in circumstances has to be internal to the grief, we have to take seriously the thought that grief is a process through which we come to terms with our loss. As Freud and Nussbaum argue, grief consists in psychological work: the detachment from the person or object we have lost. And as we complete our work, our reasons for grief diminish and perhaps eventually expire. In this sense, our change in circumstances is internal to grief, and in this sense we may be able to explain why reasons expire.

**3. Reasons Expire: Grief as Work**

In light of the view that grief consists in work, one might wonder whether I have made a mistake in framing the paradox. Perhaps we should ask, not "Do reasons expire?" but, "Do we exhaust our reasons"? Indeed, something like this idea is behind the notion of *Vergangenheitsbewältigung*, a very prominent notion in public discourse in Germany: the verb 'bewältigen' signifies something one would do with a task. *Man bewältigt eine Aufgabe*.

Nonetheless, as plausible as it may seem, there is a significant difficulty in understanding grief as work: We think of work as a goal-directed activity that aims at change. But we do not seem to think of grief as aimed at change. More precisely, we do not do so not from the deliberate perspective when we experience grief or, as I will put it, from the standpoint of grief. And this, I will argue, constitutes an insurmountable obstacle for understanding how reasons could expire.

Let me illustrate the difficulty with an example: When we work in the garden with the goal of clearing weeds from the flower beds, we continue (if all goes well) until the weeds are cleared, and then we stop. Perhaps we subsequently adopt a new goal, or we

change our focus altogether. But as we work, the thought that we will work until all weeds are cleared is not, in principle, disconcerting—though we might feel overwhelmed at the sight of an overgrown garden. There is nothing problematic as such about the thought that our work will come to an end once our task is complete. And if grief consists in psychological work, then matters should be the same with grief.

But they are not. The thought that we will grieve for a limited amount of time—for a week, or a month, or a year—is both disconcerting and distorting. To appreciate this, put yourself in the griever's shoes. Suppose you have suffered a loss. It is no comfort at all to be told that, in two months' time, or in two years' time, you will no longer grieve. Indeed, it is jarring to be told this, because it suggests that, in time, you will no longer care about your loss. Proust's narrator is acutely aware of this:

> Our dread of a future in which we must forego the sight of faces and the sound of voices which we love and from which today we derive our dearest joy, this dread, far from being dissipated, is intensified, if to the pain of such a privation we feel that there will be added what seems to us now in anticipation more painful still: not to feel it as a pain at all—to remain indifferent. (1919/2005, 340)[16]

From the standpoint of grief, the thought that we will stop grieving seems to us to amount to the thought that our loss will no longer matter to us—that, in time, we will become insensitive to it. This may be acceptable when we are upset over the end of a bad relationship or over something that we judge unworthy of grief. Perhaps a ruined shirt is like that. But the death of a loved one is different. At the moment of my mother's death, the thought that her death won't matter to me in two months' time or in two years' time is unacceptable. From the standpoint of grief, the thought that our grief will pass does not reflect an accomplishment but a failure—as a failure to care for what matters to us.

But there is a way to understand the accomplishment of grief.[17] To do so, we have to step outside of grief. We have to view it is as a condition from which we suffer. We then see grief as a process that has the functional role of getting us to come to term with

[16] Thanks to […] for drawing my attention to this passage from *Within a Budding Grove*. Moller (2007, 312) also discusses this passage. He accepts Proust's point and argues that our resilience in the face of loss is to be understood as a form of blindness to the significance of loss—a kind of delusion.
[17] My discussion here is indebted to Richard Moran's *Authority and Estrangement* (2001).

our loss—in short, a healing process. Our grief then appears to us as akin to a fever that comes and goes, as something from which we periodically suffer, perhaps especially on certain occasions, and from which we are periodically relieved. And this is not entirely inaccurate: The "pangs of grief," as DSM-5 has it, are a physiological condition that comes and goes, that decreases in intensity and that eventually subsides. To dwell on the medical analogy, we could say that grief is part of our "emotional immune system," which regulates our emotional response to our loss and from which we recover—just as we recover from a fever.[18]

The physiological view is not entirely inaccurate, because grief is a sentiment. It has an affective component which is passive in the way that our physiology is passive. Nonetheless, the physiological view is distorting, because it gives us too passive a view of grief. To the extent that we see our grief as a physiological condition—as the process through which we heal from a psychological wound—to that extent we no longer see it as our response to our reasons. A fever is not a response to our reasons; it is not *our take* on the world—even if it is the body's response to an infection. And to the extent that we see our grief as a physiological condition, or a healing process, we fail to see it as our response to our reasons.

Indeed, I think that there is a principled reason why we have no problem apprehending the temporal limitations of work but not of grief: Work, unlike grief and belief, is subject to the will. We can apprehend the temporal limitations of work, because *we* set them. But we cannot apprehend the temporal limitations of grief and belief, because we don't set them. For example, when we work in the garden with the goal of clearing weeds from the flowerbeds, we will continue, if all goes well, until the weeds are cleared. But that is because we have set out to clear the weeds from the flowerbeds: we have set that as our goal and, in so doing, we have set the endpoint of our activity. Since it is up to us to clear the weeds from the flower beds, we can decide whether to do so, when to do so, and for how long to work on this project. At any point, we can change our mind and decide that our work is done.

[18] See Gilbert *et al.* (1998) for an account of the "emotional immune system," which regulates our response to loss. I owe the reference to Moller (2007, 305).

But whether we grieve, when we grieve, and for how long we grieve is not up to us. In this respect, grief is like belief—a persistent state or activity that constitutes a response to the world, rather than a goal-directed activity that aims at change.[19] Grief is an activity or state that we can apprehend and manage—like a condition we find ourselves with. But from the standpoint of the activity or state, we do not set its temporal limitations. And that is why we cannot anticipate such limitations without alienation.

I conclude that we cannot understand grief as work, and we cannot explain why reasons expire by appeal to the accomplishments of grief. More precisely, we cannot do so from the standpoint of grief. At most, we can do so as an empirical finding about grief and the reasons for it—an empirical finding that we can grasp from a physiological point of view.

**Conclusion**

If you think about it, death is unacceptable, even if it is sometimes welcome. We cannot accept it. If we try to understand why we accept it, we become dissociated from it. We become spectators of our lives and strangers to ourselves. At the extreme, we become Meursault.

Nonetheless, we realize that, in time, we will accept all deaths, even if we don't fully come to terms with them. And this may well be a good thing for us. But that thought, however liberating it may be, is really deeply disturbing.[*]

[19] See Boyle (2011) for an illuminating account of belief as a persistent activity.

[*] Acknowledgements.

Pamela Hieronymi
UCLA

**Bio:** Pamela Hieronymi is Professor of Philosophy at UCLA. Her work addresses a variety of issues that lie at the intersection of ethics, philosophy of mind, and the theory of agency and moral responsibility. In recent years she has focused on understanding the kind of agency we exercise over our own attitudes. Her work has appeared in *The Journal of Philosophy, Philosophy and Public Affairs, Philosophy and Phenomenological Research, Pacific Philosophical Quarterly,* and elsewhere. In 2010 she was awarded the Frederick Burkhardt Residential Fellowship for Recently Tenured Scholars from the American Council of Learned Societies[1] And from 2011–2012 she was a Fellow at the Center for Advanced Study in the Behavioral Sciences at Stanford University. She is currently working on a book, *Minds that Matter*, which brings her earlier work on mental agency and notions of control to bear on the traditional problem of free will and moral responsibility.

# Standpoints and Freedom

Pamela Hieronymi, hieronymi@ucla.edu

I am presenting a piece from the second chapter of a book manuscript, and so I will begin by briefly stating where we are in the book.

The book is about free will and moral responsibility, and its first two chapters are meant to isolate what I take to be the intuitive problem of free will.

In the first chapter I present difficulties that I think do not present vexing *philosophical* problems about freedom or agency: threats to our freedom posed by interfering agents, such as meddling neuroscientists, powerful gods, and oppressive political regimes, as well as hinderances to and defects of agency, such as diseases or drugs. All of these are real threats to freedom, ones we should do our best to avoid. But, although these raise important ethical questions, and sometimes difficult philosophical questions in ethics, I suggest that they do not pose any particularly vexed philosophical difficulty about freedom or agency, itself. They are, we might say, problems in life, not in theory.

I then contrast these threats to our freedom—interferences, hinderances, and defects—with the threat that seems to be posed by deterministic physics, or mechanism. Despite the metaphorical excesses philosophers sometimes indulge, determinism is not an interfering agent—it is not analogous to a powerful god or meddling neuroscientists. It is rather a scientific claim, a claim about how the world works, one which implies that the processes that underlie and explain our agency, the processes that underlie and explain the making and executing of our decisions, unfold strictly from earlier states. But notice that the processes that underlie and explain the usual operation of our agency could not be interferences with, hinderances to, or even defects of it.

And yet, it seems, when we think about the processes that underlie and explain our agency, and when we imagine that they unfold strictly from earlier states and events, we feel our freedom is threatened—in fact, we feel we are not *really* free at all. Moreover, as noticed by many, we feel the

same intuitive threat if we imagine that our actions and decisions unfold entirely from earlier states and events in a merely lucky or probabilistic way.

This poses an especially sharp philosophical problem, in part because it is also the case that we cannot understand an event as an action, at all, unless we are able to explain it by appeal to psychological facts. To see an event as an action, we must see it as something that occurred because someone meant for something to occur. But the fact that someone meant for something to occur is a psychological fact. And, we—or, most of us—now believe that such psychological facts emerge, in their entirety, from the stuff of the earth: from nature or nurture, working in some contested combination, along with some luck. But, once we see our own actions as a part of the unfolding history of the natural world, a history that starts long before our decisions, long before even our birth, it seems to us that we are not free. And so we arrive at a vexing philosophical problem: we must see an event as explained by certain sources, to see it as an action, and we—or, most of us—think those sources are, in turn, entirely explained by prior worldly goings on. But, if our actions are entirely explained by those goings-on, we then feel we are not free—perhaps that we are not really acting, at all.

Many seek refuge in the insistence that the psychological emerges from the physical and cannot be reduced back to it.[1] But, I would argue, this fails to appreciate the strength of the intuitive problem. Shifting from neurons and chemicals to wants, desires, and beliefs, loves and commitments, fears and insecurities, self-esteem and jealousy, does not remove the worry. Loves and commitments, self-esteem and jealousy, are explained by prior states and events. Perhaps those explanations are not deterministic, but—again—probabilistic explanation is no less worrying. If some unfortunate soul, due to his or her formative circumstances, lacks the strength of ego or capacity for empathy needed to regulate his or her desires in more sociable ways, then, it seems, he or she cannot regulate his or her desires in more sociable ways. And whether she has the strength of

[1] NTS: they might thereby deny the transitivity of explanation?

ego or capacity for empathy is a matter of nature, nurture, luck, and his or her past choices. But his or her past choices are ultimately a result of nature, nurture, and luck. And, of course, just the same is true of each of us. If what you do is a function of what you are like, with or without some slippage, and what you are like is a product of what came before, with or without some luck, it seems, intuitively, that you are not free. And yet it seems undeniable that what you do is a function of what you are like, with or without some slippage, and that what you are like is a function of what comes before, with or without some luck. Thus, while it may be true that the human emerges from the physical in a way that defies reduction back to it, this will not, ultimately, assuage our concerns about our freedom. If we were bothered by Newton, Freud will do just as well.

And thus I arrive at what I take to be the intuitive problem of free will—a philosophical problem about agency. It is this: when we explain free action, we seem to explain it away. The goal of this second chapter is to try to locate the source of this problem: why should focusing on the processes or forces that underlie and explain our activities make them seem unfree or unreal?

I am not alone in thinking that explanation poses a special problem for agency—I am typically joined, in this, by contemporary neo-Kantians. The standard contemporary response to the problem is what I will call "two-standpoints" compatibilism. My task for today is to explain this response and to explain why I find it unsatisfying, both as a diagnosis of the intuitive problem and as solution to it. I will end by saying briefly where I think the real source of the problem lies.

The two standpoints in question are typically distinguished by the activities undertaken from them: There is, on the one hand, a "practical," "deliberative," "first-person," or "subjective" point of view from which we decide and act, and, on the other, a "theoretical," "explanatory," "third-person," or "objective" point of view, from which we observe, describe, and explain. This distinction in standpoints captures the intuitive problem: when we occupy the first point of view, we take ourselves to be free. But when we occupy the second, when we reflect upon our agency and

start to describe or explain it, we appear to ourselves, not as agents but as objects, and our actions appear as mere events—from this point of view, the stingy provisions of a step-motherly nature seem to curtail our possibilities. In fact, they seem to make our decisions for us. We seem to ourselves mere machines, pushed along by external determinants.

This same appeal to standpoints is also thought to ground a (to my mind peculiar) form of compatibilist response to the problem: When we occupy the second standpoint, our freedom does not appear. But, it is said, we are not entitled to conclude, from the fact that our freedom does not appear when we theorize ourselves as empirical subjects, that our freedom is only an illusion of the practical perspective. This illicit conclusion could only be reached by improperly privileging the theoretical point of view over the practical, when neither could be given priority. Even though the two points of view paint what seem to be contrasting pictures, we need not—in fact, cannot—choose between them. This is not worrying, because they concern different subject matters or conceptual schemes. The two points of view are, so to speak, *so* incompatible, that they cannot even be brought into genuine conflict. And thus we arrive at a peculiar kind of compatibilism.

I have just sketched, in bare outline, the two-standpoints approach. But notice, the outline requires filling in. *Simply* appealing to distinct "standpoints" is a compelling way to *describe* the intuitive problem. But, if we are going to do more than provide a gripping metaphor in which to state our problem, we need to know something about the two points of view—what constitutes them, why we occupy them, etc.—that might allow us to understand why they cannot be combined and so cannot genuinely conflict. And that further story might then help us understand why, when we explain our own agency, we seem to explain it away. Kant himself provided such a story, with his appeal to in-principle unknowable aspects of reality. But that is not a story that many, today, would embrace.

Notice, too, that a *mere* appeal to distinct "conceptual frameworks" or "levels of description" will not do justice to the intuitive difficulty. The intuitive problem is not the simple one that arises when we shift vocabularies or change aspects: Learning that music is explained by sound waves does not make us think music has disappeared, or that there is no such thing as "real" music, or that the music is no longer genuine. Learning that pain can be explained as neural and brain happenings has no tendency to make us think we do not *really* feel it (likewise with consciousness). In contrast, learning that (what we thought of as) agency is explicable by prior conditions makes us think there is no such thing. Our philosophical question is, Why should this be?

In a section I am cutting for time, I consider the very different labels often used, in contemporary discussion, to mark the two standpoints: "practical," "deliberative," "first-person," or "subjective," on the one hand, and "theoretical," "explanatory," "third-person," or "objective" on the other. I distinguish between (what I believe are) several distinct distinctions, and I try to show that, in each case, either the distinction does not track the apparent disappearance of agency, or, if it does, that is because we are have applied the labels by taking for granted an understanding of the intuitive problem—and so we will not illuminate the problem by appeal to a distinction between such "standpoints."

Here is one quick example from this cut section: Consider the distinction between the "theoretical" and the "deliberative" point of view. It seems ill-drawn. When I theorize about some subject matter—Newtonian mechanics, perhaps—I may well deliberate. Do I then leave the "theoretical" point of view? Surely not. So perhaps the intended distinction is really between the "theoretical" and the "practical"—the point of view of describing, explaining and understanding, on the one hand, and of decision making and acting, on the other. But, of course, in making my decisions—in deciding whether to take my umbrella, e.g., or whether to flip my omelette—I may also do some thinking about how the world works. When I do so, must I then leave the "practical" point of view, temporarily, and adopt instead the theoretical one, before returning to my practical

deliberations? If so, what, exactly, is this point of view of decision-making? Do I occupy it only at the moment of decision? When is that moment? Or perhaps I enter the practical point of view whenever I consider what people call the "normative": what is good or right or required. But surely I can theorize about such things, without making any practical decisions—and I may do so while viewing actions as entirely explained by past circumstances.

By pressing such points, I argue that, to the extent that we can draw a distinction that tracks our intuitive problem, we do by relying on our understanding of the intuitive problem. And so the distinction cannot then be put forward as a diagnosis of it, and certainly not as a solution to it.

In the end, I do not think the intuitive problem relies on a distinction between standpoints or points of view—even though it arises naturally when we reflect upon ourselves. Rather, in the end, I think the source of the intuitive problem lies in the thought (or feeling) that our own wills are not in our control, and that this thought (or feeling) arises naturally when we reflect upon ourselves, due to our confusion about what controlling our own will would require. The goal of this chapter is to arrive at that diagnosis.

However, before moving there, I would like to spend more time thinking about the "two standpoints." I think we can do a better job identifying the "two standpoints," one of which has to do with decision-making and one of which has to do with explanation, and I do think that those two "standpoints" can sometimes come into conflict. By laying out them out more precisely, and by examining more carefully how they do and do not conflict, I hope to support my claim that we will find neither the source of nor the solution to the intuitive problem here.

## THE TRUTH IN THE STANDPOINTS TALK: QUESTIONS, NOT POINTS OF VIEW

To begin, recall what I call[ed in the Introduction] *the ordinary notion of control.* When we think about what it is to exercise control, we naturally think of the control we exercise over ordinary objects, such as cars, coffee cups, and chairs, or the control we enjoy with respect to our own intentional actions, such as doing a back-flip or writing our name. These cases invite a certain model, according

to which to control a thing is to be able to conform it to your will, or, less grandly, to be able to bring the thing to be as you would have it to be. Thus it comes to seem that, in order to control a thing (your handwriting or your future), you need to have in mind how you would have it to be, and you need to be able to bring it to be as you would have it. Crudely put, exercising control of the ordinary sort is a matter of representing some change and causing the change you represent.

It is clear enough why this ordinary notion of control, with its two-part structure of controller and object controlled, leads us to think of ourselves, insofar as we are agents, as a power to effect changes in the world. Notice, though, that it also allows us, in a certain way, to ignore ourselves us as we make a decision: When you control some object (your pencil or your pan), you must have in mind the object of your control, the change you mean to effect, and (somehow) the fact that you will effect it, but you need not have in mind the psychological operations by which you exercise control. The particular features of your will that will explain your decision can remain, so to speak, behind the lens, or out of view, as you decide. And thus we introduce the visual metaphor. You are occupying what it is natural to call a "first-person," "practical," "deliberative" perspective, looking out at the world, so to speak, *from* your will, from your own point of view, rather than considering your will as though from the point of view of another. When looking out from your will, you need not have in mind any of its features.[2]

But, of course, you are not barred from considering the features of your own mind, even from your own point of view. We are sophisticated and reflective creatures, and we can think carefully about our own wills. We can sometimes understand our motives. We can often explain why we did what we did—not only by appeal to those considerations we took to count in favor of acting, but

[2] So, when thinking of ourselves as agents of ordinary control, we must think of ourselves as a power to effect change, but we may not think of ourselves as more than that. NTS: careful. you argued in chp 1 that we must consider the features of a mind to see something as an action. here you are saying we need not think of agents as having minds with features. fit these together explicitly? When you, yourself, act, you need not think of the features. Must you, to recognize your own past or future action? It seems so.

also by appeal to those features of our minds that explain why we took those considerations to so count.

But notice a relatively simple point: even if you fully understand the operation of your mind, even if you can explain your every thought and move, you cannot exercise control over your future *simply* by understanding, observing, describing, or explaining the operation of your own mind or will. To exercise control over your future, you have to make something like a decision. And, if you are going to make anything like a decision, you need to make it. No amount of observing, describing, or thinking about how the decision-making process is going to unfold will unfurl it.

## TWO ROUTES TO THE FUTURE

We need to examine this last fact more closely. Notice, first, that determining what you *shall* do, in the sense of making a decision about your future, can be distinguished from determining what you *will* do, in the sense of making a prediction about your future. You might predict that you will lose the match. This is different from deciding to throw the match.[3] Both will leave you with what is, in some sense, the same view of your future: you will lose the match. But, in the first case, you come to this view by considering ordinary evidence—considerations that show it likely that your opponent will better you. In the second case, you do so by considering, instead, features of your situation that you take to count in favor of bringing about your own loss.[4]

Likewise, you might predict—in fact, you might know—that the neuroscientists of the last chapter (who have implanted a device in your head and are able to control your thoughts remotely) are going to send you out for a walk. This, alone, will not get you walking. If you are to go for a walk *intentionally*—if the neuroscientists are to get you to walk by controlling your mind, rather than just your body—then you will have to go for a walk because you meant to; you will have to decide to go for a walk. So, if they are going to make you to walk intentionally, then they need to make you

---

[3] See G. E. M. Anscombe, *Intention* (Oxford: Blackwell Publishing Co., 1957); Stuart Hampshire, *Freedom of the Individual* (New York: Harper and Row, 1965). Cf. also CITE Hampshire, Strawson

[4] [IF you do so for reasons, you do so for such reasons]

decide. But predicting, believing, or even knowing, that you are going to make a decision is not the same as making it.

Thus there are, it seems, two routes, so to speak, to the conclusion that you will lose or that you will go for a walk: one route is occupied by predictions (in prospect, and explanations, in retrospect), while the other is occupied by decisions (in prospect, and (something like) justifications, in retrospect). You travel the first route by answering the ("theoretical") question of whether you *will* go for a walk—where that is a question you could ask about anyone (whether I will go for a walk, whether Luce will go for a walk, whether Rodney will go for a walk...). In settling this first question, you arrive at an ordinary belief, one which happens to be about yourself. The considerations you use to settle the first question (if you use any) will be those you take to show it *likely* that you (or Luce, or Rodney) will go for a walk. You travel the second route by answering a different question —not whether you will go for a walk, but, rather, whether to go for a walk. This second question is not, so to speak, about anyone,[5] and so cannot be asked about anyone else. It is, in some sense, essentially "first-personal." In settling this second question, you arrive at an intention to go for a walk. And whatever considerations you use to settle the second question will be—*in virtue* of your so using them—considerations you take, in some way, to count in favor of (or against) walking.[6]

Importantly, though, predictions and decisions routinely interact. Good decision-making often requires making predictions about yourself (whether you are likely to choke in the clutch or to forget your password). You might decide to throw the match because you predict you will lose it, anyway, and you would like to save your strength. You might decide to go for a walk because you believe the neuroscientists will make you walk and you would rather not wait around any longer.

---

[5] I am tempted to say, to answer this question is not to ascribe a predicate to a subject, and so affirm a proposition, but rather to commit yourself to effect some represented change in the world. ADD Thompson?

[6] This is not to say that you *believe* they count in favor of or against *x*-ing. It is rather that, in *using* them to settle, positively, the question of whether to *x*, by employing them in this way, you treat them as counting in favor of or against *x*-ing. AND comment about other forms of practical reasons (undermining, e.g.), and why they eventually come to bear on the practical question. THANK Wiland

Notice, again, that this entirely routine interaction of prediction and decision shows that the "standpoint" from which we make a decision can—and, in fact, ought to—avail itself of the "conceptual framework" of explanation. What is distinctive about the standpoint of decision-making is not concepts employed, but the question addressed.

So, while we might want to continue to use the metaphor of "standpoint" or "point of view," I suggest we understand the "two standpoints" by appeal to these two questions: the predictive question, of whether you will do something, and the practical question, of whether to do it. We can then consider whether addressing one kind of question, or answering it in a certain way, allows or precludes addressing the other, or answering it in a certain way.[7] Sometimes it will.

## INTERACTING QUESTIONS

Sometimes, when we make predictions, we thereby change *which* practical question we ought to address. If I realize that I simply cannot beat my opponent—if I realize that, no matter what I do or how hard I try, I will not win—then I cannot sensibly address the question of whether to win.[8] I cannot sensibly address this question because I have realized that whether I win is not up to me in the following specific sense: whether I win does not depend on my decisions, planning, skills, or effort. And thus I cannot sensibly represent winning as a change I shall bring about. And so I cannot consider whether to bring it about. I should instead adopt what I will call the *fatalistic attitude* towards winning: I should set aside the question of whether to win, and instead address some other question, such as the question of whether to do my best anyway, or to give it my all, or, maybe, to decline to compete this round.[9]

---

[7] CITE literature in formal decision theory?

[8] I *could* address it, it is *possible* for me to do so, but I would be guilty of some error. Also, CITE Bratman on the simple view

[9] [NTS: Prof P takes the fatalistic attitude toward a decision that is his to make. That's his "bad faith." The inevitability point it is confused with it. Also: another illustration of the importance of the "theoretical" to decision-making.]

It is tempting to put this point this way: the fact that my loss is inevitable makes it unreasonable for me to address the question of whether to win.[10] But this is not right. It is not the *inevitability* of my loss that renders the practical question unreasonable. It is rather the fact that whether I win does not depend on my decisions, planning, skills, or effort.

This claim needs support. To start, notice that the fact that an outcome does not depend on my decisions, etc., is, by itself, sufficient to render addressing the question of whether to bring it about unreasonable. It will do so even if the outcome is *not* inevitable. Suppose I suffer from an illness from which I might, but might not, recover. And suppose that whether I recover does not depend, in any way, on my decision-making, etc. Even though my recovery is not inevitable, I still cannot sensibly address the question of whether to recover—because my recovery does not depend on my decision-making.[11]

Even so, one might think that, if an outcome *is* inevitable, that fact, alone, makes it unreasonable to make a decision about it. But notice how odd this position turns out to be: it claims that the fact that an outcome is inevitable makes it unreasonable to make a decision about it, *even if* the inevitable outcome depends on your (admittedly inevitable) decision. This position thus declares it unreasonable to do what you will inevitably do, simply because it is inevitable that you will do it. That hardly seems reasonable—after all, the thing you will inevitably do may well be, otherwise, the most reasonable option available. (Perhaps it is inevitable that you will take the more attractive offer.

---

[10] To be precise, it is not my *loss* that is inevitable: I could simply refuse to play, and so avoid the loss. What is inevitable is, rather, that I will not beat this opponent; I will not win. For ease of exposition, I will overlook this wrinkle.

[11] As outcomes become more likely the issue becomes more difficult. There is some discussion about whether I can decide to make my free throw. CITE. I suspect this example prompts disagreement in part because learning and even accomplishing a skilled action, such as a free throw, typically involves visualizing success (repeatedly). Visualizing success and then succeeding seems (philosophically, at least) similar to representing a change and bringing it about—and so similar to deciding. But we might want to distinguish visualizing the ball going through the net from deciding to throw the ball through the net. (The two will certainly have different Bratman-style conditions and will leave one open to different questions and criticisms.) In any case, for present purposes we need not determine the point at which the unlikeliness of success requires one to change the question one addresses, if one is to remain sensible.

According to this view, the fact that it is inevitable renders the decision to take the more attractive offer unreasonable. But that seems unreasonable.)

I suspect that what underlies the temptation to think that inevitability, alone, renders decision-making unreasonable is the thought that decisions, themselves, cannot actually be inevitable—and so, if an outcome in fact depends on my decision, then it is not really inevitable, after all.[12]

But, why think decisions are never inevitable? (Or, as inevitable as anything else we take into account, when making our way through the world.) We do not generally think so, when considering other people: you may think it is inevitable that your friend will decline the offer, or investigate the misbehavior, or insult the chair. It may be said, though, that you cannot have the same view of yourself: that you cannot think your own future actions are inevitable. But, again, I think this is simply not so. It may well be inevitable that I will accept a certain job when it is offered, or tell the truth in court, or attend to the needs of my child. (As noted in the last chapter, opening a decision to contingency does not render it more free or more my own.) And, if it is inevitable that I make a certain decision, I see no bar to my knowing that.

The two-standpoints theorist can make a ready retreat to more secure ground: whether or not a decision or outcome is in fact inevitable, and whether or not I can know that about myself (in a reflective moment), I cannot sensibly *see* it as inevitable, *as* I make a decision about it. I cannot see it as inevitable, she might say, from the "standpoint" or "point of view" of decision-making.

The visual metaphor again makes the point difficult to assess, but I think there is something to this thought. After all, when you address the question of whether to walk, for example, you are, it seems, addressing the question of whether *or not* to walk. The question you decide admits of two answers: yes and no. And so it might seem that, in addressing this question, you must, in some sense, take there to be two possibilities: you could settle it positively or negatively. And, further, if you settle, positively, the question of whether to walk, you should, and you usually will, work into

[12] I am grateful to Daniela Dover for commentary on this point.

the rest of your thinking and planning the fact that you will walk, while, if you settle it negatively, you should, and usually will, work into the rest of your thinking and planning the fact that you will not walk.[13] And so, when you address the question of whether to walk, it seems you are, in some sense, contemplating two contrasting futures, each of which depends on the outcome of your decision. Thus, it might seem, to address this practical question you must treat the future as open. And thus, it might seem, so long as you continue to accept the inevitability of a given outcome, you cannot sensibly address the question of whether to bring it about.

As compelling as this last thought seems, it is not correct. First, it is not obvious that, in order to settle the question of whether to do something, you must, in any robust sense, contemplate or entertain the possibility of not doing it. Nonetheless, for the sake of argument, let us grant that, when you address the question of whether to walk, you are contemplating two contrasting futures. Let us also remind ourselves that you are also, in some way, acknowledging that which future is realized depends on your decision. In contemplating the two scenarios, you are considering whether or not to bring about some change. You have not *yet* decided the question. However, in the case we are considering, you also you believe, of yourself, that you will certainly decide the question one way rather than the other. There is no bar, it seems to me, to contemplating a future that *would* occur if you *were* to make the decision you believe you will certainly not make, nor any unreasonableness in doing so.[14] There is certainly nothing contradictory about doing so. You would, of course, be guilty of the unreasonableness already considered if you thought that the outcome will come about *regardless* of your decision, and yet proceed to make a decision about it. But you do not so regard it.

---

[13] CITE Bratman

[14] One might say it is unreasonable simply because it is a waste of time: you already know what you are going to do, so why contemplate this other possible world? Whether it is a waste of time depends, I think, on whether you know also *why* you are going to do it—whether you have your reasons for action at hand. If your secure prediction leaves your future reasons opaque to you, then you may well contemplate the two futures as you make your decision. That may be the way in which you find your reasons. But if you already have your reasons at hand, then it may be a waste of time to contemplate the future that you already believe clearly inferior. That is why I said, at the start of the paragraph, that you may not need to contemplate the alternative future.

And so, I think, contemplating the two contrasting futures while addressing the practical question can sensibly be done even while continuing to believe that one of the outcomes is inevitable.[15] Some illustrative cases have already been mentioned: I may know, in advance, that I will decide to accept a certain long-desired opportunity, tell the truth when asked, or care for my children. This does not prevent me from making my decision.

One might resist by replying that, if I *know* that I will certainly accept the opportunity or tell the truth, that is because I have *already* made the decision to do so. In any such case, my knowledge of my future action is *practical* knowledge, built on my decision. And, the opponent might continue, in advance of such practical knowledge, I cannot know what I will decide.[16]

While it may sometimes be true that I decide, far in advance, to accept the opportunity if offered or to care for my child, I doubt we must or should understand all such cases in this way. It seems possible, after all, to make predictions about your own decisions. Suppose that, while speaking to my therapist, I consider what I will do on the witness stand, and thereby come to see (what is plain to you and to him) that it is inevitable that I will tell the truth. Coming to this conclusion (or even doing so sensibly) does not require that I have *already* made the decision.[17]

If I then, later, turn to address the question of whether to tell the truth, must I, to be reasonable, expunge from my own mind what I have learned about myself? Must I suspend or revise my prediction, in order to make my decision?

---

[15] The view of the potential futures thus provides a different, and contrasting, picture, or point of view, than the view of the predictions. Here we might again want to talk about "two standpoints." But these two standpoints do not employ different concepts, and they are not in principle incompatible—in fact, it is an important part of planning to be able to reconcile them, to be able to plan to do what you see is possible. Nonetheless, in some situations they can be brought into problematic tension, as we will see below.

[16] THANK Greg for highlighting practical knowledge

[17] Perhaps I can takes steps, between now and then, to change things so that I will not: perhaps I can set up other, overwhelming incentives that I predict will motivate me to lie. My ability to do this does not undermine the relevant claim that my decision is inevitable EXPLAIN.

At just this point, the opponent may, again, appeal to standpoints: She may say, I need not suspend or revise the *prediction*, but I must, rather, enter a different point of view. From the predictive, third-person, or theoretical, point of view, I might believe that I will certainly tell the truth, but when I turn to make my decision, I adopt the practical, or first-person point of view, and I cannot continue to believe that.

But, again, I do not see what we gain by appealing to standpoints (other than some unclarity). The fact that the two *questions* are distinct, and that answering the predictive question will not, itself, amount to making a decision, is enough to do the work we need done, without restricting us further than seems real. In the case we are considering, I have settled the predictive question, but I have not yet settled the practical question: I believe (from my own point of view) both that I will certainly tell the truth and that I will tell the truth because I will decide to do so. This may be true, even if I have not *yet* decided to do so. And, as noted, no amount of predicting, nor any degree of confidence in a prediction, will simply amount to decision. And so I still have work to do. I have to get to the business of deciding. But I do not see why, in order to do *that*, I must enter anything like another "point of view" or "standpoint." I must, instead, address the practical question. And, again, I see no conflict in addressing the practical question while maintaining, "in view," my firm conviction in my prediction.[18]

To close out this point, let me consider an especially extreme case, by returning to the neuroscientists. In thinking about this case, though, it is important to remember two things. First, I am not, at the moment, wanting to make the stronger claim that inevitability is no threat to *freedom*. I am rather wanting to make a narrower point: that believed inevitability—a confident prediction—does not, by itself, render addressing and answering the practical question impossible or even unreasonable. Second, I have already granted that the meddling interferences of the neuroscientists

[18] CITE Martin Luther case. It is typically cited as a case in which I am faced with an inability on my part—so called "volitional necessity." But weaker cases will make the point I am after: inevitability is no bar to sensible decision-making.

*are* a genuine threat to freedom. But, again, I am employing them, here, to address a much narrower question: once more, whether a confident prediction will render addressing and answering the practical question impossible or unreasonable.

With those caveats in mind, recall that the neuroscientists are going to send me for an *intentional* walk—that is, they are going to control my thoughts, not just my body. Now suppose that I know the scientists will make me walk, and, further, that I have no objection to walking. Suppose I even think, all things considered, I ought to walk each day. Perhaps I have asked the scientists to make sure that I get out for a good walk today, and I know they both can and will. Perhaps they have told me they will send me walking at 9:23, and I believe them. I look at the clock. It is 9:23. I think, "Shall I go for a walk?" and answer, "Sure." And out I go.[19] In addressing the question of whether to walk—even the question of whether or not to to walk—I need not ever doubt that I will walk. And yet, I claim, I proceeded sensibly.

Here ends my attempt to support the claim that believing an outcome inevitable need not render a making a decision to bring about that outcome unreasonable—so long as the outcome depends on the decision.

It will be noticed that I have thus far focused on (what I will call) the happy cases, cases in which what I regard as inevitable and what I would have myself choose align. One might reasonably wonder: What of the unhappy cases? What if I am in the dark story in which the neuroscientists will make me decide to do something I despise? Or perhaps the Oracle tells me that the Fates have determined I will kill my father. Or maybe I simply know, of myself, that I will not follow through on my decision—as well intentioned as I am now, I will procrastinate or give into temptation. I think I should let my child cry, but I know I will not. I know I should complete the review in a

[19] I believe that I have just described the prayer life and subsequent decision-making processes of many. Other people sometimes employ life coaches. Others have elaborate strategies for ensuring that they will make certain decisions. And, when the time comes, they make the decision [and it is not true that they made the decision in advance, cf. note X].

timely way, but I know I will not. Surely, it will be said, in cases like these, my firm prediction will in some way interfere with my decision-making.

I will next consider the unhappy cases in some detail. But my final position will be this: if believed inevitability is not a problem in the happy cases, then believed inevitability does not, *in itself,* make it impossible or unreasonable either to address or to settle the practical question. Rather, I will now suggest, in the unhappy cases, certain *sources* of inevitability present hinderances to or interferences with agency. But these hinderances and interferences are, once again, problems in life not in theory. They can pose serious ethical problems, and serious problems for ethical philosophy, but they are not themselves vexed philosophical problems for agency itself. (And so, if we want to understand the intuitive problem of free will, we will again have to look elsewhere.)

In examining the unhappy cases, in which what I confidently predict I will choose is not what I would have myself choose, let us start with a very simple case, one that poses no threat to freedom at all. Suppose you are again subject to the neuroscientists, and suppose you confidently predict they will send you walking at 9:23. 9:23 comes, but you do not want to walk. You face the decision, and you think, "Nah, not right now." In that case, you will not walk (or, at least, you will not do so intentionally). The neuroscientists will have failed. In this case, the fact that you did not, at 9:23, want to walk, and so decided against walking, shows that your confident prediction was inaccurate. Such a case poses no problem for your freedom, but it also poses no threat to my claims. My claim, again, is that confident prediction about a particular outcome does not render decision-making about that outcome unreasonable, so long as the outcome depends on the decision. I have also been assuming that decisions, like any other event in the world, can be inevitable. But I need not assume either that the neuroscientists are omnipotent or that you are infallible in your confident predictions. So, it may be that they, or you, get it wrong sometimes.

Let us turn, next, to the familiar example of Professor Procrastinate. Procrastinate is asked to review an article within a given time frame. To agree to complete a review on time is to commit to a

plan of action that will require a number of other decisions along the way.[20] And we can assume that, if Procrastinate made the right subsequent decisions—if he put away his book, avoided making another cup of coffee, spent less time surfing the internet, etc.—he would certainly succeed. So, unlike our ill-fated competitor, whether Procrastinate succeeds depends (and, we can suppose, entirely depends) on his own decision-making. His difficulty is that he can predict (and, we are granting, can accurately predict) that he will not make the required decisions, when the time comes. How should Procrastinate proceed? What decisions can he sensibly make?

First, and importantly, notice that, unlike in the case of the ill-fated competitor, it seems sensible for Procrastinate to *address* the question of whether to complete the review—in fact it seems he is *required* to address that question and return an answer to the editor. Whereas the ill-fated competitor could not sensibly address the question of whether to win, since winning did not depend on her decisions, etc., Procrastinate *must* address it—because the outcome does depend on his decision. But Procrastinate knows as well as we do that he will not write the review, even if he accepts the invitation. So how is Procrastinate to answer the question he must address?

Like any of us, Procrastinate cannot sensibly agree to write the review unless he can be reasonably confident that will do it. Most of us are entitled to that confidence, without first drawing up elaborate plans: we can reasonably count on ourselves to sort it out as we go. But, given Procrastinate's poor record on such things, he cannot sensibly proceed in this way. If he did so, then, in light of his track record, he would be guilty of bad planning. Nor can he sensibly decide to rely on strategies that have failed in the past. And so Procrastinate needs to have in mind some reasonably detailed plan, in which he can have confidence. Notice, though: if he has such a plan, and if his plan is tolerably reasonable, then, it seems, it is no longer inevitable that he will fail, and he can sensibly accept the request.[21]

---

[20] CITE Michael Bratman's pioneering work.

[21] [CITE Beri's work about the "reasonable"]

But what if Procrastinate finds himself unable to come up with any reasonable plan? What if he continues to regard it as inevitable that he will fail? He cannot, then, sensibly agree to complete the review. Can he sensibly decide to decline it?

It seems problematic, in some way, for Procrastinate to decline the review because he regretfully predicts that he will not complete it, if, as we have stipulated, whether he completes the review depends entirely on his own decision-making, planning, efforts, etc. It seems to be in some way in bad faith. In fact, I think there are at least two different problems, what might be thought of as two different kinds of bad faith, to be distinguished.

First, Procrastinate would be in bad faith if he treated the prediction, *itself*, so to speak, as settling the practical question. As we have noted several times now, the predictive question and the practical question are distinct, and answering one does not simply amount to answering the other. So, your confident prediction that you *will* walk is not yet a decision *to* walk, and Procrastinate's confident prediction that he will not complete the review is not yet a decision to decline it. One form of bad faith—what, it seems to me, Sartre had in mind—would try to ignore this distinction and allow the prediction just to stand in, so to speak, for the decision. But this will not do. If you walk intentionally, you will walk because you mean to walk—and so you will need to settle for yourself, positively, the question of whether to walk. Likewise, if Procrastinate declines the review, he will have to settle, positively, the question of whether to decline it. He cannot avoid that decision.

So a prediction cannot simply stand in for a decision. Nor does a prediction, simply by itself, bear on a decision. But you might take it so to bear. You might take your confident prediction that you will walk to bear on the question of whether to walk.[22] Perhaps you are impatient and would

[22] You will then take it, in some or another (perhaps indirect) way, either to count in favor of or to count against walking Why will I take it to so bear, at least in some indirect way? Because any consideration that I employ, in answering the question of whether to act, must ultimately, in some or another way, come to bear on that question, and so ultimately, in some or another way, be taken to either count in favor of or count against walking. Thanks to Eric Wiland for comments on this point.

rather avoid waiting around: might as well get this over with. Or maybe you think walking, now, will somehow help you to retain some sense of control over your own future. Or maybe, since resistance is futile, you would simply like to save your strength. Alternatively, maybe you think it is a reason not to walk—since it shows the walking unfree—but, in the end, other needs won out. So while a prediction cannot simply stand in for a decision, it can become one consideration (typically one among many) in light of which you decide.

Just the same is true for Procrastinate. His prediction, that he will not complete the review, cannot simply stand in for his decision. But it is relatively easy to see how he might take the prediction to count in favor of declining the review. He might reason, "I am sure not to complete the review, so it will be best for everyone if I decline now." In so deciding, Procrastinate has not simply treated his prediction as if it were a decision. He is making a decision, one for which he can be held, and hold himself, to account. So a charge of one form of bad faith [the form that, I think, goes with the charge of self-deception] will not stick.

But his decision can still seem problematic. He is still deciding not to complete the review because he foresees he will not, and he takes the fact that he will not do it to count in favor of not doing it. What is the remaining problem?

Here is one very tempting way of answering (a way that I used to endorse, but now think is mistaken): The question Procrastinate is addressing is the question of whether he *shall* complete the review, and he cannot, in addressing that question, treat as given the fact that he *will not* complete it —because whether he will complete it is precisely the matter under consideration. Likewise, the question he is addressing (whether to complete the review) takes into consideration the subsequent decisions required to do so—in deciding whether to complete the review, Procrastinate is also, therein, deciding whether to do what is necessary to complete it (whether to set aside the required time, whether to get up early, etc.). And so he cannot, when addressing this question, treat as given the fact that he will not do what is required. That is, again, precisely what is under consideration.

As tempting as this response is, I now think it is not right—again, because of the happy cases. In those cases, there is no difficulty with taking your own future decisions to be inevitable—in fact, there is no difficulty in taking the inevitability of your future decision to act to count in favor of deciding so to act. I may be confident that I will care for and attend to my children, and that confident prediction may be part of my reason for deciding to adopt children; or I may be confident that I will relentlessly pursue justice, and that confident prediction may be part of my reason for accepting a certain challenging job. So it seems to me that the difficulty with Procrastinate cannot be that he treats his future decisions as inevitable, nor even that he takes the inevitability of his future decisions into account when making them. Rather, the problem, I suggest, is simply that he, himself, regards the future decisions he plans to make as poor ones, even as he plans to make them. Or, rather, to put the point in a cleaner way [23]: the problem is that he takes the fatalistic attitude towards his own future decisions and starts to plan around them.

But why is that a problem? Return to the thought that, when you decide on some course of action, you are contemplating a future. You are also committing to a plan, a plan that might include a range of sub-decisions.[24] As you make the decision to complete the larger task or project, you are also, therein, committing to make those needed decisions along the way (that is why, in light of his past failures, Procrastinate cannot agree to complete the review without a reasonable plan). So, in committing to his plan, Procrastinate is committing to make the required decisions along the way. Those decisions are, then, in some sense, included in the decision he is making now. So if Procrastinate, in deciding to decline, is treating his future procrastinating decisions as facts to plan around, he is in some way treating those decisions as though they are not up to him. But, we have agreed, they are up to him. So he is in some way incoherent. It is as though he is counting on something or someone other than he—or, other than the he now making the decision to decline—

---

[23] NTS: "poor" might drag you into guise of the good. But that should be needless.

[24] CITE Bratman.

to shore up the decision to procrastinate, when the time comes. Something or someone else must, so to speak, hold those future decisions in place, or explain them, to make sense of his decision-making now. And that other something else, whatever it is, is a threat to his freedom, in the sense that it is an *interference with*, *hinderance to*, *constraint on*, or *defect in*, the operation of his agency.

What, then, is a procrastinator to do? In the end, I think that, if Procrastinate really cannot find a plan that would enable him to have any confidence that he will complete the review, then he ought to engage in this lesser kind of bad faith—he ought to plan around his own regrettable decisions, fatalistically, in the same way that the dominated opponent must plan around her loss. But, whereas the dominated opponent works around her loss because it does not depend on her own decision-making, Procrastinate is working around what is, everyone agrees, up to him. And so he is treating his future his self as though he were another (and unreasonable) person. (It is clear why it is tempting to call this taking up a "third-person" point of view on yourself. It is also clear why it might seem an evasion of responsibility—how are we now to hold you responsible for these future decisions you now disavow?) It is a bad position to be in. But the problem is not exactly one of self-deception or inauthenticity (as it would be, if he pretended that the prediction settles the matter). Procrastinate may be vividly and accurately aware of his predicament, and he may be doing everything he can to figure out to do, to take responsibility for it. It is rather a problem of disunity —and it is a *defect* of agency. This is not a problem posed simply by the inevitability of a future decision.

Much the same can be said, and, I think, in the case of the Fates and the Evil Neuroscientists, where our hero faces, not a defect of his own agency, but rather external manipulation and interference. [I will skip these, for time, luckily].

## CONCLUSION

So, what have we accomplished? Our question was: why does explaining agency seem to explain it away? Why does agency seem unreal, once we learn that it is natural? Some have thought that we

can explain this by pointing out that the "standpoint" of explanation and the "standpoint" of decision-making are not only distinct, but so different as to make it the case that they cannot be brought into conflict.

I have suggested that the distinction between standpoints is best understood as a distinction between questions: between a predictive question and a practical question. I have insisted that answering one question does not amount to answering the other, and I have suggested (though I have not directly argued) that this fact is all we need to understand the apparent difference in "standpoint" or "point of view." I have examined how and why answering the predictive question with certainty can make it unreasonable to address or to answer the practical question. Unreasonableness appeared only in what I called the unhappy cases. In these cases, the "point of view" of prediction and the "point of view" of decision do part ways. But these are also, I have argued, cases in which you see your agency as subject to some hinderance, interference, or defect.

Thus we have not, I think, yet found anything to help us understand why, when we explain our own free actions, we seem to explain them away. [We could think we were just confused, but this is not satisfying]. And so I think our original problem remains. I believe I can say what it is. I think it is not, in the end, a problem with standpoints or points of view, however rich and important these ideas are. It is, rather, a somewhat simple problem about our ordinary notion of control—our ordinary notion of control will not allow us to see our own will, our own decision-making (or concluding, believing, caring) as in our control. But if our decision-making, concluding, believing, and caring are not in our control, then it seems that nothing really is. That is the problem I propose to address in the coming chapters.