ABSTRACT

Title of Dissertation: NAVIGATING MORALITY, AGENCY, AND EMOTION IN INTERPERSONAL CONTEXTS

Lia Rachel Curtis Fine, Doctor of Philosophy, 2026

Dissertation directed by: Distinguished University Professor Peter Carruthers, Department of Philosophy

This dissertation seeks to address certain issues that are central to moral psychology broadly construed. At its core are questions about interpersonal harms, our practices of blaming, and the process of moral repair. In the first chapter, "Gaslighting and Self-Deception," I draw a novel connection between the phenomena of gaslighting and self-deception. Specifically, I argue that the victim of gaslighting finds herself in a state of "imposed," or "induced," self-deception of the Melean variety. I argue that in addition to the epistemic harms normally associated with gaslighting, this constitutes an additional harm to the victim. Not only is the victim of gaslighting left in an epistemically impoverished position, but the added layer of "imposed" self-deception creates an additional harm to both her agency and her autonomy. In the second chapter, "Blame and Norm Psychology," I discuss our practices of blaming and holding individuals responsible. I argue that despite there being many extant

theories of blame, none of them capture all the social and moral aspects and functions of blame. Theories that frame blame as a communicative device have trouble accounting for blame that goes unexpressed. Theories that suppose blame is a costly social signal have trouble explaining detached blame. Most relevant theories struggle with some aspect or form of blame, yet each extant theory also seems to get something right. I propose that we turn to the evolutionary discipline of norm psychology to make sense of these diverse phenomena. Norm psychology posits mechanisms that create mutually re-enforcing and stable cooperative groups, and includes those based on reputation, punishment, signaling, aspects of cultural transmission, cooperative disengagement, and combinations of these. I argue that blame is one such candidate mechanism. Finally, in the third chapter, "Acceptable Apologies," I address the issue of moral repair. I argue, in this chapter, for a distinction between grades of supererogation: I intend to distinguish between nodes on a continuum of moral repair. I argue that there is a morally salient difference between the acceptance of an apology and the forgiveness of a wrongdoer, and while both are supererogatory acts, when faced with a sincere apology that sincerely promises changed behavior, it might be a "morally permissible moral mistake," to refuse to accept. However, the same cannot be said of forgiveness. The incorporation of acceptance as an intermediary point between the rejection of an apology and the forgiveness of the wrongdoer allows for: (1) the completion of the wrongdoer's apology wherein the process of restoration can begin, and (2) the victim to have more cognitive space to rationally process her resentment, and to do so without an obligation to forgive before the restorative process occurs.

NAVIGATING MORALITY, AGENCY, AND EMOTION IN INTERPERSONAL CONTEXTS

by

Lia Rachel Curtis Fine

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2026

Advisory Committee:
Professor Peter Carruthers, Chair
Professor Hallie Liberto
Professor Dan Moller
Professor Elizabeth Schechter
Professor Melanie Killen

Dedication

For Lucille, Chad, and Ross; in loving memory.

Acknowledgements

I am deeply indebted to so many individuals for their support and encouragement in the writing and preparation of this dissertation. First, I want to extend special thanks to my advisor Peter Carruthers. Peter is the best advisor I could have asked for, and were it not for his unwavering support, I could not have written this dissertation.

Many thanks are also due to everyone on my committee, their input has helped me grow immensely a philosopher and a scholar. Numerous conversations with Elizabeth Schechter and Dan Moller have shown me not just how to refine my ideas, but what good philosophy looks like, and all the various forms it can take. Special mentions are due, however, to Hallie Liberto. Hallie's influence has really shown me how high of a standard I can hold myself to, and that such a standard really is achievable. My ideas are better because of all of them.

Thanks also to all the faculty and colleagues in the philosophy department and beyond at the University of Maryland, I am a better philosopher and understand the field better because of everyone here. Thanks especially to; Harjit Bhogal, Sam Kerstein, Paolo Santorio, Fabrizio Cariani, Melanie Killen, Rachel Singpurwalla, Nick Laskowski, Eric Pacuit, and Dan Blair. Thanks for always letting me come knock on your doors and chat about whatever topic I have on my mind. And very special thanks to Louise Gilman, nothing would get done without her.

Of course, I would be nowhere without the support of my friends and the philosophical community of graduate students at UMD. Special thanks to my various writing group peers over the years: Joe Gurrola, Shen Pan, Robert Ragsdale, and Xiaohui Yu – all of your feedback on my work has been invaluable. Thanks also to the entire grad student community here but especially to: Aidyn Cooper, Jeremiah Tillman, Arundhati Chakraborty, Jackson Iszler, Gavin Victor, Ben

Fan, and Xu Jiang. Our endless conversations and debates have greatly shaped who I am as a philosopher.

I also want to thank my friends and particular faculty from my MA at Brandeis University. Without them, I would not have even made it to the PhD. Special thanks to Beri and Jen Marušić, and Jerry Samet. Thanks also to: Louis Doulas, Dan Friedman, Kellan Head, Casey Rufener, Ripley Stroud, and Evan Welchance.

I want to extend special thanks to my undergraduate philosophy department at UNC-Asheville. The department is now, very sadly, dissolved, but its impact will never be forgotten. Very special thanks to: Keya Maitra, Duane Davis, Brian Butler, Grace Campbell, Melissa Burchard, Scott Williams, and Gordon Wilson. Additionally, extra special thanks to particular faculty from the UNC-Asheville psychology department as well: Patrick Foo, Michael Neelon, and Elizabeth Pascoe. Without all of your influence I would not be the interdisciplinary philosopher I am today.

Personal thanks are due to my many friends for helping me stay sane, especially Farah Aboul-Nasr, Ian Anderson, Travis Boyd, Bailey Bruce, Monika Chao, Dylan Greer, Simon Kapiamba, Nora Leonard, Isabelle Parris, Connor Rattey, Sonya Robinson, Joey Small, Vanessa Singh, and AJ Wright. Thank you for your constant encouragement and support, for traversing the worlds of gaming and film with me, for our many late night (and occasionally early morning) conversations, and for always being there when I really needed someone.

Of course, the most important acknowledgements come last. I'd like to thank my parents, who taught me how to argue, and how to think. They always assured me that I could do whatever I wanted in life, and I am eternally grateful to them. And lastly, I'd like to thank my partner,

Christopher Masciari. Thank you for always, *always* encouraging me to keep going, and for always being there for me when the going got really tough.

Table of Contents

Chapter 1: Introduction

PROSPERO:
Hast thou, which art but air, a touch, a feeling
Of their afflictions, and shall not myself,
One of their kind, that relish all as sharply
Passions as they, be kindlier moved than thou art?
Though with their high wrongs I am struck to th' quick,
Yet with my nobler reason 'gainst my fury
Do I take part. The rarer action is
In virtue than in vengeance. They being penitent,
The sole drift of my purpose doth extend
Not a frown further. Go, release them, Ariel.
My charms I'll break, their senses I'll restore,
And they shall be themselves.
-The Tempest (V.I 21-32)

This dissertation aims to address certain issues that are central to moral psychology, broadly construed, and to our interpersonal moral lives. At its core are questions about interpersonal harms, our practices of blaming, and the processes of moral repair. The epigraph from *The Tempest* above captures these concerns in a particularly vivid way. Prospero, having been wronged by his brother Antonio, orchestrates a shipwreck that brings his enemies within his control. Through magic, he manipulates their perceptions and induces in them fear, guilt, and confusion, subjecting them to suffering in response to their wrongdoing. His aim is not *merely* to retaliate, but to bring them to recognize their offenses and to feel the moral weight of them. Yet, even as he is "struck to the quick," by their betrayal, Prospero ultimately refrains from further vengeance. Influenced in part by Ariel's expression of compassion, and in part by his own reflection, he resolves instead to forgive them, to restore their senses, and to allow them to "be themselves."

What Shakespeare captures in this moment is a central feature of moral psychology: the tension between affective response and reflective restraint in the face of wrongdoing. Prospero's

response is not that of a purely ideal rational agent, rather, it reflects the non-ideal, affectively charged, messy, and often morally ambiguous terrain of interpersonal moral life. It is precisely this terrain that I seek to examine in this dissertation. Specifically, I focus on instances of harm, our practices of blaming, and the processes by which moral repair may (or may not) be achieved.

Each of the following chapters engages with distinct but overlapping areas of moral psychology, including work on moral harms (e.g., gaslighting), blame, and moral repair (e.g., apologies and forgiveness). Rather than treating these as isolated phenomena, I examine how they arise within the broader context of interpersonal moral life. The primary goal of this introduction, then, is to situate and motivate each of my three projects and my approach towards them.

1.1 Underexplored Moral Harms (Gaslighting)

"You're overreacting." "That's not what happened." "You're being paranoid." "That's crazy." Even for those unfamiliar with the term gaslighting, these kinds of utterances may feel immediately recognizable. They are not merely instances of disagreement or dismissal of dissenting views. Rather, they are characteristic of a more insidious interpersonal dynamic in which one individual is systematically led to doubt her own perceptions, memories, and beliefs. In recent philosophical and psychological literature, this phenomenon has come to be known as gaslighting.

Standard accounts of gaslighting emphasize its epistemic dimension and the impoverished epistemic state that the victim finds herself in. On these views, the central harm lies in the way the victim's epistemic standing is undermined: her confidence in her own epistemic capacities is eroded, often to the point that she comes to see herself as unreliable, irrational, or "crazy" (Abramson 2014, 2024). While this captures something important, I argue in my first paper,

"Gaslighting and Self-Deception," that it does not fully account for what is going wrong in cases of gaslighting.

My central claim is that gaslighting involves not only epistemic harm, but a distinctive form of "imposed" or "baited" self-deception. Drawing on work by Mele, Funkhouser, and Trivers, I argue that victims of gaslighting are not *merely* deceived by others, but are recruited — without full awareness or control — into patterns of motivationally biased reasoning that sustain false beliefs about themselves, their environment, and most prominently, the gaslighter. In this way, the victim becomes, in part, an unwitting participant in her own epistemic undoing.

This additional layer is philosophically significant because it reveals a deeper moral harm. If gaslighting involves the systematic induction of self-deception, then it does not merely undermine the victim's epistemic standing; it also disrupts her autonomy and agency. By shaping the very processes through which she interprets evidence and forms beliefs, the gaslighter compromises the conditions under which the victim can think, deliberate, and act as an independent agent.

1.2 Blaming Practices & Normative Psychology:

Imagine that you have asked a friend for help moving house. They agree, promise to be there, and then, when the time comes, they simply fail to show up — leaving you high and dry to manage on your own. In response, you might reasonably feel anger or resentment; you might reconsider the relationship; and you might, quite naturally, *blame* them.

Blame is a pervasive feature of our moral and social lives. We blame in many different ways and for many different reasons: sometimes directly (to the party that has wronged us), sometimes to third parties, and sometimes only privately, "in our own hearts." Our blaming can be affective or detached; it can be expressed or remain entirely unspoken. We can blame ourselves;

we can blame the dead; we can blame in myriad ways. This diversity makes blame a concept which any good theory needs to do some "fancy dancing,"[1] to accommodate. Any adequate account must explain not only why we blame, but also why it takes such varied forms.

Existing philosophical theories of blame each capture part of this picture. Communicative accounts emphasize the role of blame in expressing disapproval and initiating moral repair; protest accounts highlight its function in marking wrongdoing; and signaling theories understand blame as a costly display of one's normative commitments. While each of these approaches identifies an important feature of our blaming practices, none is able to accommodate the full range of cases. In particular, they struggle to explain forms of blame that are unexpressed, self-directed, emotionally detached, or directed away from the wrongdoer.

In "Blame and Normative Psychology," I argue that we can better understand this diversity by situating blame within the framework of norm psychology. On this view, blame is not best understood as having a single function, but as a multifunctional mechanism embedded within a broader system of norm articulation, enforcement, and maintenance. Blaming practices help regulate behavior, coordinate expectations, manage reputations, and stabilize cooperation within social groups.

This perspective allows us to unify the insights of existing theories without reducing blame to any one of them. What communicative, protest-based, attribution-based, and signaling accounts each capture are not competing explanations, but different aspects of a single underlying mechanism. By understanding blame in this way, we gain a more comprehensive account of why it is such a stable and pervasive feature of our moral psychology.

[1] Phrase borrowed from Shoemaker & Vargas, 2019.

1.3 Moral Repair (Apologies & Forgiveness):

"I accept your apology, but I'm not ready to forgive." We are all familiar with moments like this. A wrongdoer apologizes, perhaps sincerely, and yet something remains unsettled. One might accept the apology — acknowledge the effort, even appreciate it — while still feeling resentment, still unable to forgive. Indeed, it is a common, but unrecognized, feature of our moral lives that acceptance of a good apology, and forgiveness of a wrongdoer can come apart.

This ordinary distinction, however, sits uneasily with much of the philosophical literature. Many accounts of apology and forgiveness treat the acceptance of a sincere apology as intimately connected with (if not sufficient for) forgiveness. On such views, a successful apology provides strong reason, and sometimes even an obligation, for the victim to forswear resentment. Yet this does not align with our day-to-day experience. We frequently accept apologies *without* forgiving, and we sometimes forgive in the absence of any apology at all.

This tension raises a cluster of questions. What makes an apology good, or rather, acceptable? What, exactly, is accomplished when an apology is accepted? And how, if at all, is this distinct from forgiveness? In "Acceptable Apologies," I argue that answering these questions requires introducing a distinction that is largely absent from the literature: the distinction between acceptance and forgiveness as two distinct points on a continuum of moral repair.

My central claim is that acceptance and forgiveness occupy different grades of supererogation. Forgiveness, I argue, is always supererogatory: no matter how sincere or complete an apology may be, a victim cannot incur an obligation to forgive. Acceptance, by contrast, occupies a more complex normative space. In cases of sincere, "promissory apologies" — those that commit the wrongdoer to better future behavior — it may be the morally right thing to accept,

even though refusal remains permissible. In such cases, refusing to accept may constitute what Elizabeth Harman calls a "morally permissible moral mistake."

To develop this view, I first argue that many interpersonal apologies are best understood as promissory acts that generate obligations for the wrongdoer. I then articulate a set of conditions under which an apology can count as acceptable, including sincerity, performativity, and what I call an allocution condition, which requires that the wrongdoer demonstrate an understanding of the wrongdoing. Finally, I show how distinguishing acceptance from forgiveness allows us to make sense of cases in which moral repair begins without requiring the victim to relinquish resentment. By introducing acceptance as an intermediary stage in moral repair, we gain a more nuanced account of our interpersonal practices—one that preserves the supererogatory nature of forgiveness while explaining how restoration can begin even in its absence.

Chapter 2: Gaslighting and Self-Deception

1. Introduction

In the extant literature on gaslighting the harm done is often thought to be an epistemic one. The victim is undermined in her own trust of and confidence in her memories, her perceptions, and her beliefs. Kate Abramson (2014, 2024) defines gaslighting as, "a form of emotional manipulation in which the gaslighter tries (consciously or not) to induce in someone the sense that her reactions, perceptions, memories and/or beliefs are not just mistaken, but utterly without grounds—paradigmatically, so unfounded as to qualify as crazy" (Abramson, 2014). It is clear that this constitutes an injustice to the victim. Exemplar cases of gaslighting clearly result in epistemic injustice: because of the manipulation of the gaslighter, the victim comes to no longer take herself seriously, or be taken seriously by others, as a source of knowledge, or as an independent deliberative perspective (Abramson, 2014). Indeed, in Abramson's terms, the victim comes to believe herself to be "crazy," or "unreliable," or in the cases of moral gaslighting that Kate Manne (2025) describes, "a bad person." Really, the victim comes to believe about herself whatever it is that the gaslighter wants her to believe. What makes this so morally wrong is not just that the victim's epistemic standing is undermined which is of course a very bad thing, but also that the beliefs that the victim comes to hold about herself or about the world or about the gaslighter as a result of being gaslit are false. Of course, undermining the victim's epistemic capacities is a serious wrongdoing. But what's worse is that in gaslighting, the victim must also accept these false beliefs. Indeed, the victim herself is recruited – without full awareness or control – by the gaslighter to aid in her own undoing. In her oft cited book The Gaslight Effect (2007), Robin Stern even argues that it is necessary that the victim play an active role in the cycle of gaslighting. Indeed, she calls gaslighting "a tango," because, after all, "it takes two,"

(Stern, 2007). How does this happen, exactly? Well, by being taken advantage of and manipulated in particular ways such that the victim adopts certain false beliefs about herself, the world, or the gaslighter—specifically the false beliefs that the gasligther really wants the victim to come to accept. Perhaps surprisingly, the active role that the victim plays in being gaslight is quite important to the conversation here, but it is not the whole story. In the following, I will argue that nearly all varieties of gaslighting—epistemic (Abramson 2014, 2025), moral (Manne 2025), and even perhaps structural/institutional (Berenstain, 2025; Longhair, 2025)—the victim is brought into a state of so-called, "baited self-deception," (Funkhouser, 2019), or "imposed self-deception," (Trivers 2000, 2002). Further, I argue that the flavor of self-deception that the victim is "baited," into, is essentially that which Mele expounds, e.g., a state of motivationally biased reasoning. This matters for the following two reasons: 1) that the victim is brought into a state of "imposed," self-deception is a descriptively accurate outcome of gaslighting that has not yet been identified in the literature, and 2) this constitutes an additional harm to the victim; in being "baited," into self-deception she is not merely harmed in her epistemic standing, but also in her autonomy and her agency.

In what follows, I will argue that in being gaslit, the victim is brought into a state of "baited," or "imposed," self-deception, and that this is additionally harmful. In the first section, I will discuss the phenomenon of gaslighting in general. I will give examples and discuss what gaslighting consists in, some different varieties of the phenomenon, and why it is the case that some people may be more vulnerable to falling victim to it than others. In the second section, I discuss a particular view of self-deception; specifically, Mele's deflationary account, which resolves many puzzles and paradoxes of the traditional views of self-deception, and fits the gaslighting cases quite nicely. Then, I give an explanation of Funkhouser's (2019) concept of

"baited self-deception," and of Trivers' "imposed self-deception" (2000/2002, 2011). I also show how all three views hang together. In the third section, I will bring these two literatures into discussion to illustrate how the "baited" or "imposed" self-deception, described by Funkhouser (2019) and Trivers (2000/2002, 2011), is essentially what occurs in gaslighting cases. In section four, I will make the moral point: I will argue that if it is wrong to manipulate someone, and to deceive someone, then to recruit someone by gaslighting them into their own self-deception is clearly doubly wrong. Finally, at the end, I conclude.

2. Gaslighting

While the term gaslighting has recently become popularized, the concept remains somewhat difficult to pin down. However, the term is more popular than ever before, and it does seem that people widely misuse it. It's become an all-too-common colloquialism that many people use to, say, spurn mere disagreement (e.g., someone raises an objection to an argument, and the person making the argument responds, "don't gaslight me!"). While that is hardly my topic here, it is good to first say what gaslighting is not. Gaslighting is not just disagreement about the facts in some situation, it is not the same as mere lying or deception (though lying and deception are generally involved in gaslighting), and it is not the same as mere dismissal of someone's views. There is a great deal of confusion and fuzziness around the concept itself and this is a big part of why nailing down exactly what is going on in the phenomenon of gaslighting remains elusive, even though, as mentioned, the term is more widely used than ever before. In this section, I discuss what gaslighting is, various forms it can take, and who is most vulnerable to it.

2.1 What is Gaslighting?

So, what is gaslighting? It is an interpersonal form of abuse or emotional manipulation where one individual, the gaslighter, attempts to severely reduce the victim's confidence in her own perceptions, memories, and beliefs. Gaslighting is always interpersonal, and it quite often occurs in relationships that are particularly intimate (e.g., romantic relationships), or where there is some power dynamic at hand. In "A Theoretical Framework for Studying Gaslighting," Klein et al. (2025), note that "[…] close others shape and verify our self-views and experience of the world." Kate Abramson, in her 2014 paper, "Turning Up the Lights on Gaslighting," defines the term nicely: gaslighting, "[…] is a form of emotional manipulation in which the gaslighter tries (*consciously or not*) to induce in someone the sense that her reactions, perceptions, memories and/or beliefs are not just mistaken, but utterly without grounds—paradigmatically, so unfounded as to qualify as crazy" (Abramson, 2014. Emphasis added). Named after the eponymous 1944 movie "Gaslight," gaslighting is a generally interpersonal phenomenon, and a form of manipulation. Gaslighting necessarily occurs across numerous instances over time, via multiple interactions (Klein et al., 2025), where one party (the gaslighter) systematically brings the other party (the victim) to lose trust and confidence in her own perceptions, memories, and beliefs. Let's look at a few examples for clarity:

> ***Gaslight* (1944):**
>
> In the movie "Gaslight," Gregory, the gaslighter, is after his wife Paula's family jewels. He knows that the jewels are hidden somewhere in their attic, but he needs a reason to go there, or he needs to go there in secret. Every night he tells Paula he is going to work, but instead he sneaks to their attic where he then searches for the jewels. However, every time he goes into the attic, he turns on the gas lamps there which causes them to dim in the other parts of the house. Every time this happens, Paula asks him about it, and every time she asks, he tells her that it isn't happening. While Gregory's changing of the lights in the rest of the house is unintentional, when asked about it, he denies that they have thus changed. And he does so with the conscious intention of undermining Paula's confidence and trust in her own perceptions, thus "driving her crazy," such that he can ship her off to an asylum

(and finally steal the jewels and her wealth for himself). Eventually, Paula figures out what Gregory is up to, with the help of a local policeman. However, during the time that she is successfully gaslit, Gregory tells her time and time again that she must be mentally unwell, so much so that eventually she comes to believe it herself, at least in part.

Relationship Advice Forum:

A post on a relationship advice forum wherein an anonymous woman is seeking advice reads:

"We've been fighting a lot lately because of conflicting work schedules. […] Recently though things have been happening in our apartment that makes [sic] me feel a little crazy, I've been getting hurt a lot, and Alex [the boyfriend] keeps reprimanding me that I need to be more careful and not be so clumsy, but honestly it feels like it's not me, but that things are being done/moved/placed?

About three days ago I came home from work and was preparing something to eat in the kitchen […] I opened up one of our kitchen cupboards and the heavy door flew off one of the hinges and smacked me in the head. I have a huge goose egg and a giant bruise near my hairline, enough that several co-workers have asked me about what happened. […] But here's the thing, as I tried to put the door back together - I couldn't find the hinge or three screws anywhere afterward, they should've been on the floor or the counter, and they weren't.

[Today] I opened up Alex's underwear drawer to get a pair of socks to wear to bed, there were the screws, the hinge, and a screwdriver. […] I feel like my apartment is booby trapped, and maybe my boyfriend did it on purpose to hurt me, but I feel insane bringing it up."[2]

***Dirty John* (as quoted by Kate Manne, 2025):**

Dirty John is a podcast produced by the *LA Times*. It follows a woman, Debra Newell, who falls in love with and marries, within just months of meeting him, a conman John Meehan. "He pretended to be an anesthesiologist (dressing up in scrubs on their dates), while in reality he was a nurse anesthetist who had been fired for stealing drugs intended for patients […]." Among other things, John had recently been released from prison for felony drug theft just before meeting Debra, a fact which he hid from her. In fact, the prison employees thought he was so deceptive and awful, that they gave him the nicknames: Dirty John, or alternatively Filthy John. "Eventually, [Debra] found incontrovertible evidence of his myriad

[2] Note: this is one excerpted example from a very long post that does reflect a pattern of multiple instances of this type over a longer period of time. Read the full post here: https://www.reddit.com/r/BestofRedditorUpdates/comments/1ivb2gl/i_27_f_think_my_boyfriend_29_m_boobytrapped_our/

deceptions—arrest warrants, prison records—and moved out of their shared home […]. When Debra withdrew from him, he began to threaten her and depicted her as the wrongdoer: accusing her of stealing from him, hitting him, and other supposed misdeeds she never committed. […] Nonetheless, somehow, despite all this Debra not only forgave John but was persuaded by him that it was all a big misunderstanding—she bought his demonstrable, dangerous lies [over and over again]" (Manne, 2025).

The following is an excerpt from an interview from the podcast with Debra and an *LA Times Journalist:*

Debra: He was trying to tell me so many times that he was set up and had to go to jail. Please forgive him. He just knew I wouldn't understand until he had all the evidence in front of him.

Journalist: All a big misunderstanding?

Debra: All a big misunderstanding and he had an answer for everything; and it was so convincing that I thought, Okay. He, literally, had convinced me, at this point, that he is not this person.

Journalist: Despite all of the paperwork?

Debra: Yes. All the facts were right there in front of me and he is that convincing that I would say that […] I was also in love with him. It's so hard, when you're in love, to listen. You're listening with your heart, not your head.

Journalist: Did you ask about his nickname, Dirty John?

Debra: He said it wasn't true. He said, "I don't know where you got that from." It was as if everything […] He was able to convince me. He was so good at it, it could be a cold day out and he could convince me that it's 95 degrees, that's how good he was. To where you questioned yourself.

Journalist: It's almost like he convinced you that all the facts about his life were some kind of hallucination on your part?

Debra: Yes, he made me out to be the one […] That he was this great guy and that everyone else had done him wrong, is what he said […] [H]e always, again, he always had a story. He told me that he had lied because he thought he'd lose me, that he feels so lucky that I'm such a forgiving person who, hell, I'm the love of his life, that I've made him a better person. Just all this kind of stuff […] I felt guilty, to some degree […][3]

[3] Listen to the whole *Dirty John* podcast here: https://www.latimes.com/projects/la-me-dirty-john/

Liz the Executive (Stern, 2007):

> "Liz is a top-level executive in a major advertising firm. A stylish woman in her late forties with a solid twenty-year marriage and no children, she's worked hard to get where she is, pouring all her extra energy into her career. Now she seems to be on the verge of reaching her goal, in line to take over the company's New York office.
>
> Then, at the last minute, someone else is brought in to take the job. Liz swallows her pride and offers to give him all the help she can. At first, the new boss seems charming and appreciative. But soon Liz starts to notice that she's being left out of important decisions and not invited to major meetings. She hears rumors that clients are being told she doesn't want to work with them anymore and has recommended that they speak to her new boss instead. When she complains to her colleagues, they look at her in bewilderment. "But he always praises you to the skies," they insist. "Why would he say such nice things if he was out to get you?"
>
> Finally, Liz confronts her boss, who has a plausible explanation for every incident. "Look," he says kindly at the end of the meeting. "I think you're being way too sensitive about all this—maybe even a little paranoid. Would you like a few days off to destress?"
>
> Liz feels completely disabled. She *knows* she's being sabotaged—but why is she the only one who thinks so?

These are just a few examples of how gaslighting can function interpersonally. While each example is contextually different, the common features are there. For the most part, for my purposes here, I will focus on romantic relationships, as I believe these are exemplar cases of gaslighting. However, as the last example shows, gaslighting can occur in many contexts. In each case, there is a "close-other" of some kind, be it a partner, or a boss (you could imagine other cases too—parent and child, or a manipulative friend). And in every case, it seems that the victim desires the approval of the gaslighter. Each case results with the victim questioning her capacities or coming to believe that she herself is "crazy," or "stupid," or "a bad person." There are some common utterances that the gaslighter often uses: "you're overreacting," "you're too sensitive," "that's crazy," "you wouldn't do that to me," or "that didn't happen the way you thought." Indeed,

even if one is not aware of the phenomenon as such, many of these utterances might feel familiar. Generally, when gaslighting is discussed in philosophical contexts, it is carved out as a grave moral wrongdoing because of its epistemic aspects; the undermining of the victim's beliefs, perceptual capacities, and reliability of her memory. Indeed, gaslighting is clearly a moral wrong because it results in things like silencing (as described by Langton, 1993; and expounded by Abramson, 2014), and epistemic injustice (Fricker, 2007). As we will see, it is not *just* these effects of gaslighting that are at issue here. But also that the beliefs that the victim comes to hold about herself are false, and she enters into a variety of "baited" or "imposed" self-deception as a result of being gaslit.

In the eponymous case, *Gaslight* (1944), during the time that she is successfully gaslit, Paula really does start to believe she is mentally unwell. Gregory does other things to induce this belief in her. The changing of the gas lamps is just one among many things that he does which slowly cause her to lose her mind. For instance, at one point in the film, Gregory gives Paula a brooch that he claims was his grandmothers, and thus it is very valuable to him. He gives it to Paula and tells her he wants her to wear it, but then immediately takes it back, stating that it may be slightly broken, and he must get it fixed before she can wear it because in its current state it could easily fall off of her and get lost. Then, at a later point in the movie, Gregory asks Paula where the brooch went, and when she says she doesn't have it, he accuses her of losing it. He says to her, "Remember? I gave it to you! Oh, you've gone and lost it, haven't you! You're always losing things! I think your memory is going! Perhaps you are unwell!" At first, she swears she doesn't have it, she *remembers* that he took it back, but eventually with Gregory telling her over and over that she must have lost it, she accepts his explanation: she must have lost it, she really is forgetful.

The second example, from the relationship advice forum, is actually quite close to the original usage of the term. In this example, the gaslighter is changing aspects of the victim's physical surroundings (e.g., their home), and then denying that they themselves had anything to do with the occurrences. The result again in this case is that the victim feels that she is going crazy, and in this particular case it seems she is brought into a state of general confusion. She feels her home is booby trapped and she doesn't understand what is going on.

In Manne's case of *Dirty John*, Debra is brought to believe that *she* is the bad person, when it is clear that her conman husband John is the immoral one. In the excerpt from the interview with the *LA Times* journalist, Debra admits to being completely convinced by John, despite all of the evidence, the paperwork, and the facts that she knew about the situation. She still came to accept that what she knew to be true couldn't be possibly be the case. Despite the evidence in front of her, she was still, in some way or another, convinced.

And finally in the case of Liz the Executive (Stern, 2007), her career is at stake, and so she wants her boss to think well of her. She starts to doubt her own perceptions, and tries to adopt the boss's (Stern, 2007). But as Stern notes about this case, "[…] her boss's view of things really doesn't make sense to Liz. If he's *not* trying to sabotage her, why is she missing all those meetings? Why are her clients failing to return her calls? Why is she feeling so worried and confused? […] Wishing desperately for her boss to be right, but knowing deep down that he isn't, makes Liz feel completely disoriented, no longer sure of what she sees or what she knows" (Stern, 2007). Liz is very clearly gaslit by her boss, and resultantly she is also brough into some variety of self-deception.

So, it seems that there are commonalities among these disparate examples. There are certain properties that all gaslighting shares. As Klein et al., (2025) concisely put it, there are two

core features of gaslighting: "(1) an attempt by the perpetrator to convince a target that they (the target) are epistemically incompetent […] and (2) the target's epistemic trust in the perpetrator." This is quite important. As we will see, it is because of (2), that the victim ends up in the ultimate state of self-deception as a result of being gaslit. If the victim did not trust the gaslighter, and did not give the gaslighter's explanations, and opinions a higher than ordinary level of credence, then she would not even be a candidate for gaslighting.

2.2 Species of Gaslighting:

While gaslighting often occurs in romantic and intimate relationships, it can also occur in other contexts and domains. In this section, I will discuss some of the other domains in which gaslighting often occurs, and some other varieties of gaslighting. As I have already discussed interpersonal romantic gaslighting, I will not discuss that further in this subsection. However, just like in the exemplar cases of gaslighting in romantic relationships, in these other cases as well the gaslighting must occur over time, the gaslighter must make the victim feel epistemically incompetent in some way, and the victim has a level of epistemic trust in the gasligther and desire for their approval such that she eventually accepts the gaslighter's perspective. For the purposes of space, I will only survey a few relevant species of gaslighting in this subsection.

2.3 Moral Gaslighting:

In her chapter, "Moral Gaslighting," Kate Manne (2025), describes a slightly different, but I would imagine, quite common variety of gaslighting. Moral gaslighting, similar to garden variety gaslighting, involves the gaslighter undermining the victim's self-evaluation of her own capacities. Though in this case, the capacities that are undermined are not epistemic, but moral. Manne defines moral gaslighting as cases, "[…] in which someone is made to feel morally defective – for

example, cruelly unforgiving or overly suspicious—for harboring some mental state to which she is entitled" (Manne, 2025). Manne uses *Dirty John* as an example of moral gaslighting. As we saw above, whenever Debra, the wife, would start to question John or pull away from him, he would accuse her of being a bad person in some way or another. It was whenever she would begin to catch onto his schemes or get close to the truth that he would lash out and attempt to further manipulate her. But Manne is correct that in this case, the gaslighting wasn't about Debra's epistemic capacities; it was about her moral capacities. And because Debra wanted to be a certain type of person (e.g., a "good" wife, trusting, kind, forgiving…) she was susceptible to being manipulated in this way. Manne even argues that this phenomenon, moral gaslighting, may be easier and more effective, "[…] than impugning a targe's rationality to gaslight her" (Manne, 2025). Essentially, in general, we don't often doubt that our perceptions and even our values are grounded in reality. It would take a lot for someone to convince most of us that we, say, hallucinated. However, it might be much easier to convince a morally conscientious agent that she has enacted some moral failure. It seems clear that moral gaslighting is another legitimate species of the phenomenon.

2.4 Who is Vulnerable to Being Gaslit:

There are lots of different forms that gaslighting can take. But it seems that, in most cases, the gaslighter and victim are in an intimate relationship. Indeed, as Klein et al (2025) notes, the fact that the gaslighter is a "close other," to the victim matters a lot. Klein et al., (2025) show that close others shape our self-view according to the relational self view (Klein et al. 2025; Andersen and Chen, 2002). In articulating the relational self view, Andersen and Chen (2002), argue that the self is an associative network of relationships, consisting of linkages between

representations of significant others and representations of the version of the self we are with each significant other (Andersen and Chen, 2002; Klein et al. 2025). Essentially our significant or close others inform our self-perceptions, and our informational processing about the world. Specifically, social influence, a basic feature of social psychology, plays a powerful role in how we view both ourselves (as mentioned above), and the world. "Shared reality," is a specific type of social influence which is particularly important and informative in cases of gaslighting. Klein et al (2025) explain that, "Experiences of shared reality can occur in many kinds of relationships (with friends, acquaintances, and even with strangers), they are especially powerful in romantic relationships because shared reality acts as a means of verifying one's beliefs about the world while also facilitating feelings of interpersonal closeness, thereby satisfying both epistemic and relational needs" (Klein et al. 2025). So, it is precisely because the gaslighter is in a privileged position, because they are a "close other," to the victim that they can exploit features like the victim's epistemic trust. And it is similarly the victim's closeness to the gaslighter (or her love for them, or, trust, or admiration, or her deference to them if, say, in a professional setting) that makes her vulnerable to being manipulated in this way. Importantly, it isn't enough that just anyone is merely denying certain facts about reality; this is a denial of some important fact by a close other that the victim trusts. Surely if a random stranger on the street came to one of these victims of gaslighting and said, "Did you know? Birds are fish!" They would surely respond, "What are you talking about?" But, because the gaslighter has special standing to the victim, because they are a close other whom the victim trusts, she is more vulnerable to be manipulated. Indeed, similarly, Klein et al. (2025), explain, "[…] imagine someone approaches you and says, "aliens have landed on earth!" The degree to which this statement causes you to update your view of reality will be greater if the person is a close other rather than an acquaintance or

stranger" (Klein et al. 2025). Indeed, in Gaslight (1944), Paula is willing to entertain that she really could have kept and lost the brooch. The woman in the advice post is willing to consider that maybe the door really did just fall off the hinges like that. Debra in Dirty John, is willing to consider that maybe she is the one who is wrong. And Liz the Executive (Stern, 2007), is willing to think that maybe she really is just being paranoid. But it also seems to be the case that in all these examples the victim, "knows, deep down, that their gaslighters are telling them something that doesn't ring true" (Stern, 2007).

Different individuals will be differently positioned to be more or less vulnerable to the different forms of gaslighting. For instance, individuals who have been historically marginalized are more vulnerable to be structurally gaslit, than people who are in positions of power (Longhair, 2025). Perhaps it is just generally true, that for any form of gaslighting, individuals with histories of mental health issues are more vulnerable than the average individual, as they might be more likely to doubt themselves than someone who doesn't have a history of mental illness. In a similar way, women may be more vulnerable to being gaslight, in general, than men because of gender inequalities and the fact that women may possess a similar propensity to second-guess themselves more than men do. It is certainly the case that many feminist theorists argue that this is true (Gunn, Longhair, & Oliver, 2025). All of this being said, however, this does not mean that anyone could not fall prey to this kind of manipulation given the right circumstances, power dynamics, and relational properties.

Section 3: Self-Deception:

Self-deception is classically modelled on interpersonal deception. Much like interpersonal cases, it is said that the self-deceived must simultaneously hold a dual belief: p and $\sim p$. This, of course gives rise to many paradoxes, for instance, the *static paradox* which addresses the law of

non-contradiction: how can one simultaneously believe two contradictory beliefs ($p + \sim p$)? Indeed, the possibility of self-deception, "[…] seems to pose an impossible state of mind […]" (Deweese-Boyd, 2023). In this section, I will discuss self-deception insofar as it can shed light on what is happening in cases of gaslighting. I will explain how one version of self-deception may work. In order to avoid puzzles and paradoxes, I will hone in on a particular definition of self-deception that is also most apt for my purposes here; specifically, Alfred Mele's deflationary account of self-deception as non-intentional motivationally biased reasoning. Additionally, in this section I will explain Trivers' (2000/2002; 2011) view of "imposed" self-deception, and Funkhouser's (2019) "baited" self-deception.

3.1: Melean Self-Deception:

In his works "Real Self-Deception," (1997) and Self-Deception Unmasked (2001), and elsewhere, Mele gives a deflationary account of self-deception that avoids certain paradoxes and puzzles traditional accounts are fraught with. Rather than attempting to account for what appears to be an impossible mental state that traditional accounts must explain (e.g., that someone could simultaneously believe both p and ~p), Mele instead argues that one enters into a state of self-deception when one is sufficiently motivationally biased by some desire to treat evidence in a particular way (Mele, 1997). This can be cashed out in various ways, like: confirmation bias, selective attention, positive or negative misinterpretation, and selective gathering of evidence (Mele, 1997). Mele argues, "A plausible hypothesis about [these tendencies] is that our wanting something to be true sometimes exerts a biasing influence on what we believe" (Mele, 1997). Importantly, Mele also argues that we needn't enter into such a state of motivationally biased belief intentionally. Indeed, he argues that we generally do such things unintentionally. Mele states, "Such strategies of self-deception as positive and negative misinterpretation, selective

attending, and selective evidence-gathering do not depend for their effectiveness upon agents' employing them with the intention of deceiving themselves" (Mele, 1997). And this is surely even the case if one conceives of intentions as mental states like beliefs or desires. Though, to this point, Mele does defend a view of intentions as "[…] executive attitudes toward plans, in a technical sense of "plan," that, in the limiting case, treats an agent's mental representation of a prospective "basic" action like raising his arm as the plan-component of an intention to raise his arm" (Mele, 1997, 1992).

In any case, it seems that Mele's motivationally biased self-deception is the most plausible story for such cases. This explanation also fits neatly with the aforementioned gaslighting cases. However, as we will see, there is one extra feature in gaslighting cases: that the self-deception is "imposed," or "baited" by a "close other."

3.2: Imposed & Baited Self-Deception:

Both Funkhouser (2019), and Trivers (2000/2002; 2011), give accounts of self-deception that can be induced by another individual. This might seem odd, and the natural question that arises is: why is this not merely garden-variety deception? But it is precisely *not* that, because in these cases some other individual either imposes some (implicit or explicit) threat, trigger, motivation, or incentive for the agent to deceive themselves.

In Trivers' cases of "imposed self-deception," others provide us with a motivation or incentive to self-deceive (Trivers 2000/2002). Trivers work on self-deception is largely influenced by his work as an evolutionary biologist, and he gives a corresponding story: parents could be incentivized to shape their children to have cognitive and behavioral tendencies that further the parents' interests and the child's interests—for instance, putting the collective good of the family above those of the individual child, which may induce self-deception regarding the value of

altruism (Trivers, 2000/2002). But, additionally, and more relevant for my purposes, Trivers also discusses how self-deception can be imposed by putting an individual in a situation that compels self-deception as a coping mechanism. Trivers, in his 2011 book, *The Folly of Fools* […], argues, "We are highly sensitive to others, and to their opinions, desires, and actions. More to the point, they can manipulate and dominate us. This can result in self-deception being imposed on us by others (with varying degrees of force)" (Trivers, 2011). He gives quite a few examples here: a captive may come to identify with her captor, an abused wife may take on the worldview of her abuser, and a molested child may blame herself for the transgressions against her (Trivers, 2011). As well as functioning as a coping mechanism, Trivers posits that functionally, for the victims, self-deception in such cases may also be for the purpose of reducing conflict with the dominant individual (Trivers, 2011). Trivers even goes so far as to explain that this is often the view that many self-deceived victims take *themselves*, "An abused wife may be deeply frightened and may rationalize acquiescence as the path least likely to provoke additional severe assaults—this is most effective if actually believed" (Trivers, 2011). It seems quite plausible that such "imposed" self-deception is possible, and perhaps occurs in a great many cases.

Similarly, Funkhouser (2019), posits another variety of self-deception that is initiated by others, what he calls "baited self-deception." "Baited self-deception," is much like the Trivers' "imposed self-deception," but in these cases, "[…] the victim has a standing motive or tendency to believe something, and others act so as to exploit that vulnerability. The person is baited or lured into a self-deception, but the person already had the motive and then does most of the heavy lifting once the motive has been triggered and perhaps directed by another" (Funkhouser, 2019).

Importantly, the self-deception in the baited and imposed cases needn't be the traditional self-deception where the agent-patient is lying to themselves and believes p and $\sim p$. In fact, it

seems much more plausible that in these cases of self-deception initiated by others, the self-deception is far more likely to be of the variety described by Mele (motivationally biased, unintentional, etc.). Indeed, it does seem that Trivers is describing something at least approximating Mele's view—Trivers' self-deception (imposed or otherwise) need not require the presence of conflicting beliefs. And similarly, Funkhouser's "baited self-deception," can also easily be captured by Mele's conception as well. As we will see, all of these views together give us an interesting picture of what may be happening in cases of gaslighting.

Section 4: Gaslighting & Self-Deception:

The connection between gaslighting and self-deception is hopefully becoming clearer. The picture I want to paint is something like this: in gaslighting, the victim is brought to believe that she is epistemically incompetent by the gaslighter. However, while the victim eventually accepts the gaslighter's worldview or beliefs about her (the victim's) competence, she seems to also exhibit some cognitive dissonance, even as she accepts such beliefs, because these beliefs are not true. And so, as a result, the victim enters into a Mele-style form of self-deception. She may positively or negatively misinterpret events or actions of the gaslighter, to see them in a certain light. She may engage in confirmation bias, looking for evidence that supports the claims the gaslighter wants her to support in order to placate him, or win his approval. She may selectively attend to certain aspects of the gaslighter or their relationship so as to avoid the reality of the situation. Or she may selectively gather evidence about the gaslighter or the situation to attempt to maintain the beliefs that the gaslighter wants her to accept. Of course, all of this may occur without the victim intending to do any of it. But by accepting the beliefs or the worldview that the gaslighter wants her to accept, the victim is in so doing, engaging in a form of "baited" or "imposed" self-deception.

Section 5: The Moral Upshot:

5.1: Additional Moral Harms:

Now, why does any of this matter, morally? We know that gaslighting is morally wrong in one sense because it results in an epistemic injustice to the victim. However, there is another harm that I want to carve out here. In being gaslit, and in being induced into a state of self-deception, the victim is not only harmed epistemically, but she is additionally harmed in her autonomy, and thus also in her agency. There are a few things at work here. One is our conception of the self, which will be critical in a theory of autonomy. Earlier, I articulated the relational view of the self (Andersen & Chen, 2002; Klein et al., 2025). This view is going to inform the theory of autonomy we must also endorse. I do not have the space to articulate an entire theory of autonomy here, however, a quick sketch is helpful. Endorsing the relational view of the self, would seemingly entail something like "relational autonomy" (Stoljar, 2024). Relational autonomy is a feminist reconception of autonomy that articulates autonomy as essentially interpersonal and intersubjective, and critically dependent on our close social relationships. In her article, Stoljar states, "If relationships of care and interdependence are valuable and morally significant (cf. Mackenzie & Stoljar 2000b, 8–10), then any theory of autonomy must be "relational" in the sense that it must acknowledge that autonomy is compatible with the agent standing in and valuing significant family and other social relationships" (Stoljar, 2024). So, this view of the self and of autonomy combined, will have specific moral implications for what is going on in cases of self-deception via gaslighting.

It seems like if our conception of autonomy on the one hand, and our self-view and worldview on the other, are importantly linked to close others and relationships of care (Stoljar, 2024; Klein et al., 2025), then being recruited into a state of "baited" self-deception, or "imposed"

self-deception by a significant "close other," would surely constitute a disruption of autonomy, minimally. In a more robust sense, this kind of induced self-deception may even disrupt the victim's agency, and deliberative capacities. If someone is brought into a state of self-deception, as Mele describes it (e.g., a state of motivationally biased reasoning), they will not be able to appropriately gather evidence, accurately interpret events, test "hypotheses," (meant in a broad sense), or attend to the most relevant features of a situation. This is surely a disruption to their deliberative capacities, and thus their autonomy and agency. For instance, if someone is self-deceived in the Mele sense, and they desperately want to believe that their spouse is faithful when they are not, they will treat weak evidence for faithfulness as strong; they will treat strong evidence of *unfaithfulness* as ambiguous; they may fully avoid looking for further evidence or confirmatory information; and they may misinterpret suspicious events as innocent. If autonomy, and relatedly, agency, require that the individual's actions are grounded in an appropriate appraisal of reasons, *and* in our close relationships with others, then someone who is in a state of imposed self-deception caused by gaslighting can certainly not be said to be autonomous. If the victim's deliberation is structured around motivationally biased reasoning, her choices may feel free and autonomous, but they are actually not so.

5.2: Moral Responsibility, Reconsidered:

One might worry that describing the victim of gaslighting as engaged in a form of self-deception risks attributing to her a kind of responsibility for her epistemic state. If self-deception involves motivationally biased reasoning (e.g., selective attention, confirmation bias, and so on,) then it might seem that the victim is, in some sense, responsible for her own deception. This would be a deeply problematic outcome. However, this worry rests on an outdated picture of self-deception.

On the Melean account of self-deception that I endorse above, self-deception does not require intention, nor does it involve the agent knowingly sustaining a false belief. Rather, it is the product of ordinary cognitive mechanisms operating under motivational pressure. As Neil Levy argues in his 2004 paper "Self-Deception and Moral Responsibility," once we abandon the traditional view of self-deception as intentional and paradoxical, we must also abandon the presumption that self-deceivers are thereby morally responsible. Responsibility for belief requires a form of control, specifically, the ability to recognize and correct for biasing influences. But, in paradigmatic cases of Melean self-deception, agents lack this kind of awareness.

This point applies with particular force in the cases of gaslighting that I outline here. Not only is the type of self-deception involved of the Melean variety, and so it already avoids many problems or moral responsibility, but it is also *imposed* self-deception. The victim's epistemic environment is systematically manipulated by a trusted "close other," over an extended period of time. As a result, the very conditions require for responsible belief formation (i.e., awareness of doubt, recognition of bias, and the ability to critically reassess one's beliefs,) are undermined. The victim does not simply fail to correct for bias. Rather, she is placed in a situation in which such correction is not reasonably available to her.

Therefore, characterizing the victim as being in a state of imposed self-deception does not imply that she is blameworthy. On the contrary, it helps to explain the depth of the harm: the victim is not merely deceived by the "close other," but is recruited into cognitive processes that she cannot adequately monitor or control. The resultant imposed self-deception is therefore not a failure of responsibility, but a manifestation of the injustice she suffers.

5.3: Structural, Institutional, and Medical Gaslighting:

In recent years many feminist theorists have discussed what has now been dubbed "structural gaslighting," or "institutional," gaslighting. In her chapter, "Theorizing Structural Gaslighting […]" Holly Longhair gives a clear explanation of this species of the phenomenon. In these cases there is no one identifiable agent who is the gaslighter, "[…] but instead a system holding in place the conditions that gaslight the members of particular groups" (Longhair, 2025). Similarly, the features that are constituitive of gaslighting can be easily applied to contexts of phenomena like institutional betrayal (e.g., betrayal by a religious leader, or an academic department chair, etc.), or a therapeutic alliance of a doctor and patient. As Klein et al. (2025) note, "Epistemic trust is a core feature of many of these alternative relationship contexts—indeed […] authority figures (doctors, religious leaders) are particularly relevant as these individuals are thought to "know best," which could establish a set of priors that facilitate gaslighting. In these cases, there may be one single individual (the gaslighter) who leverages the victim's trust in the institution that the gaslighter represents.[4]

Section 6: Conclusion:

In this paper I have articulated what gaslighting is, which is no easy feat. I have shown that gaslighting cannot be mere disagreement about facts, it cannot be mere lying or deception, and it is not purely one-sided. I have shown that gaslighting involves the active participation of both the gaslighter and the victim, and that in being gaslit, the victim eventually is brought to accept the worldview or beliefs that the gaslighter wants her to accept. In so doing, the victim is

[4] There are lots of other varieties of gaslighting expounded in the literature, though for reasons of space I will not spend time on them here. But other notable varieties of gaslighting are affective gaslighting (Oliver, 2025), epistemic gaslighting (Ivy, 2025), and manipulative gaslighting (Stark, 2025).

"baited," into a state of self-deception by the gaslighter. The particular flavor of self-deception that the victim finds herself in is essentially that which Mele describes in his work: a state of motivationally biased reasoning. Further still, I hope to have shown that if a victim is induced into such a state of self-deception via gaslighting that this is not just a harm to her epistemic standing, but also to her autonomy.

Chapter 3: Blame and Normative Psychology

1: Introduction:

Imagine a scenario: you are scheduled to move house and have asked for my help. I agree and promise I will help you move when the time comes. Then, the time arrives for me to help you move and I blow off my commitment, leaving you on your own to move your belongings. There are many responses that one might have to such an event. You may feel anger and resent me for leaving you high and dry. You may think to yourself that you will no longer ask me for help on such endeavors. And most importantly for our purposes here, you may *blame* me for my failure to hold up my end of the deal. What is the social function of a reaction like blame, and why do we so naturally feel the impulse to blame those who have done wrong? Quite a common response to acts like broken promises, direct harms, and wrongdoings, blame is a robust feature of our psychology, our social interactions, and our moral ecology. In ordinary experience we blame in many different and disparate ways. We engage in direct blame of wrongdoing targeted at oneself[5], and we also engage in indirect third-party blame. We blame in ways that are affective and emotional, and we blame in ways that are detached and cold. We express blame to others, and we blame privately in our hearts. We blame the dead, we blame ourselves, and we blame each other. So, it seems that any good theory of blame has a lot to account for.

Indeed, extant philosophical theories of blame all have to do some "fancy dancing,"[6] to accommodate some or other of these phenomena. For instance, theories that frame blame as a communicative device have trouble accounting for blame that goes unexpressed. Theories that suppose blame is a costly social signal have trouble explaining detached (or non-affective) blame.

[5] Most argue that this constitutes paradigmatic blame.
[6] Phrase borrowed from Shoemaker & Vargas, 2019

Most relevant theories struggle with some aspect or form of blame, yet each extant theory also seems to get something right about blame. Communicative theories of blame are correct in asserting that when blame is overt and dyadic, it does function to communicate something to the blamee. Costly signal theories of blame are right in showing that blame functions as a signal to both the blamee and present third parties, for instance. It is just because blame is so multifaceted and multifunctional, that as it stands no one theory can adequately explain all of blame's disparate functions and expressions.

To make sense of these diverse phenomena, I suggest we look to "norm psychology" and its evolutionary basis. Humans are hyper social creatures. We often socialize with individuals outside of our immediate families and, in fact, we exhibit widespread cooperation among non-relatives. In such socializing conflict is bound to occur be it conflict amongst individuals, intra-group conflict, or inter-group conflict. As such, there must be some stabilizing and organizing mechanisms that make such large-scale socialization possible and mediate the inevitable conflict. Norms are one such crucial organizing mechanism that help with the stabilization of social groups. Norm psychology is the claim that we have evolved psychological mechanisms and dispositions to adopt norms, internalize them, use them to guide our behavior, and enforce them in various ways. An evolutionary approach, norm psychology illustrates not only our psychological capacity for norms but also how we acquire them, how they interact, how they are enforced and upheld in social groups, and how they support the cooperation of individuals in social groups. Norm-psychology lends explanation to questions like the above: how is cooperation sustained in groups of unrelated individuals? But also, how do we cooperate when we are continuously interacting with new individuals? And, how do we keep social groups together and functioning? Norm-psychology posits models and mechanisms that result from adaptive selection pressures and

ultimately keep social groups functional and cohesive. Mechanisms that create mutually re-enforcing and stable cooperative groups include those based on reputation, punishment, signaling, aspects of cultural transmission, cooperative disengagement, and combinations of these (Chudek & Henrich, 2011). I will argue that blame incorporates all these functions. A norm psychological account of blame can unify blame's disparate theories and give us an extended and all-encompassing account.

In the following, I will argue that endorsing a norm psychological account of blame will give us a more robust theory of blame, and one which can account for all of the varieties of blame that cause our extant philosophical theories of blame to come up lacking. In the first section I discuss blame. In this section I explain what blame is and how it can manifest differently in different contexts. In the second section I discuss the disparate functional theories and the varied functions of blame. In this section my aim is to show that while all of these theories seem to get some aspect of blame right, they all struggle to explain certain features or aspects of blaming. In the third section, I discuss the norm psychological account and how this account gives us an extended and unifying account of blame that incorporates even more functional features. In this section I also discuss precisely what such an account of blame entails and means. I conclude that while most extant theories of blame fall short in some way or another, an evolutionary approach motivates a form of norm psychology that seems to give us an all-encompassing account.

Before proceeding, a clarification about the ambition of the present project is in order. In proposing a unifying explanation of blame, I do not assume that all extant theories are attempting to capture every phenomenon that goes by the name "blame." Some accounts focus on paradigmatic interpersonal blame; others adopt a more pluralistic framework that distinguishes between different kinds of blaming practices. My aim is not to deny the legitimacy of these

approaches, but to ask whether there is a deeper explanatory thread that runs through the many phenomena we call blame.

Relatedly, some theorists may deny that certain putative cases—such as purely private moral appraisal—count as genuine instances of blame at all. The argument that follows does not depend on every marginal or controversial case qualifying as blame in the strictest sense. Rather, I aim to show that many of the central and widely discussed varieties of blame can be illuminated by situating them within the broader framework of norm psychology. Even if some edge cases resist this framework, demonstrating that it explains a wide range of paradigmatic and contested cases would already constitute theoretical progress.

The unifying ambition of this paper is therefore modest but significant: I make no commitment to what really does or does not fit our shared concept of blame. This project is not in the business of conceptual analysis. Rather, I aim to show that there are a diverse range of phenomena that can be explained and unified from a norm-psychological perspective, each of the instances of which some philosophers, at least, have called instances of blame. Thus they may all be understood as emanating from the same source whether or not they are all correctly called "blame," as the term is normally understood.

The contribution of my project then, is not to replace existing accounts of blame, but to provide a unifying explanatory framework. By situating blaming practices within the broader context of norm psychology, I aim to explain why blame exhibits both the evaluative features emphasized by appraisal-based accounts, and the social-functional features emphasized by communicative, protest-based, and signaling theories.

Section 2: Blame:

What is it to blame someone? While I could devote much space to the conceptual analysis of the concept of blame, I will leave that discussion for another time. What I aim to do here is to simply note the different forms that blame can take. One thing that generally all the leading theories of blame agree on is this: "blame is a response to a person in light of his or her perceived norm violation, where the blamer takes that violated norm seriously" (Shoemaker & Vargas 2019). So, we blame when we see that some norm that we have internalized has been broken, and we take the breaking of such norms seriously. The most paradigmatic cases of blame are what have been called "direct overt blame" in some cases, and "dyadic blame" in some other instances (Shoemaker & Vargas 2019). This simply means that interpersonal blame is generally considered the paradigmatic case of blame. So, paradigmatic blame is occurrent when some agent calls out some other agent for some wrongdoing or norm violation directly to the offending agent.[7] This is to be distinguished from other types of blame like third party blame where the blamer blames the wrongdoer to some other third party, or private blame which would go unexpressed. So, blame can come in different varieties and can differ along a few dimensions. Most importantly for our purposes here, blame can be direct or third-party oriented, affective or detached, and expressed or unexpressed.

There are certain features shared by nearly all varieties of blame. Most basically, blame is both social and cognitive, and blame regulates social behavior. In our paradigmatic cases of blame we have one individual who has been wronged, and one individual who has enacted some harm or wrongdoing by breaking some norm: the blamer and the blamee. There might be many things the

[7] Note that I say some wrongdoing *or* some norm violation because one could blame someone for violating a nonmoral norm. We can blame others for violating many types of norms be they moral, social, sartorial, etiquette norms, or others.

blamer wishes to accomplish in such instances. Perhaps most foundationally, in blaming the blamer is attempting to draw out some response from the blamee, and in turn to regulate their behavior. The blamer may desire a promise of better behavior from the blamee, or an apology, or some attempt at restoration of the relationship or the harm they have caused. In addition to wanting some response from the blamee, the blamer might blame in order to signal to the blamee and to any present third parties something about themselves: that they do care about this norm and its violation, and they are willing to police such breaches. In other types of cases the function is less clear: what about cases of blame that go unexpressed? In cases of unexpressed blame, the blamer might be motivated by their blame internally, which may cause them to cooperatively disengage with the blamee. In the following, I discuss the different dimensions of blame and why these differences matter for our broader discussion.

2.1: Affective and Detached Blame

No matter the variety of blame, be it direct and overt, or directed at a third party, etc., our blame may be either affective or detached. In her paper, "Responsibility Without Blame […]" Hannah Pickard argues for an account of blame that is effective in the clinical treatment of individuals with cluster B Personality Disorders. While I will not be focusing on the clinical or psychiatric aspects of her theory here, I will focus on a critical distinction that she makes in this paper. That is, the distinction between affective blame and detached blame. Pickard argues that detached blame, not affective blame, is clinically effective in the treatment of these particular disorders. Now what does this distinction amount to? Affective blame, for Pickard, is blame that carries it's "characteristic sting." (Pickard 2011). This is to mean that judgements and expressions of affective blame are not merely cognitive but are also emotional. For Pickard, affective blame can consist in a number of emotions, "most obviously, these include hate, anger, and resentment,"

but these emotions can also include, "[…] disapproval, dislike, disappointment, indignation, and contempt" (Pickard 2011). So, in affective blame, the emotional component of blaming is wrapped up with the cognitive component as well. Detached blame on the other hand is defined by Pickard as blame without its characteristic "sting:"

> Detached blame can consist in a judgement or belief of blameworthiness. It can be accompanied by a revision of attitudes or intentions, or a further belief that such revision would be appropriate. It can also be accompanied by the imposition of negative consequences for the action, or just a demand for accountability or answerability. The point is that it need not have any of blame's characteristic 'sting' (Pickard 2011).

Now this is just to say that detached blame is possible and that it might look like a judgement of blameworthiness. While it is possible to enact detached blame, our impulse in blaming tends to be affective in nature. In Pickard's clinical contexts, the clinicians are modulating their affective impulses in blaming, and thus, blame in a detached manner. But such modulation may not be easy. It might even be the case that the affective component of blame is built in for many but not all instances of blame, and any detached blame is more effortful in its expression. For the rest of us in non-clinical contexts our impulse may be the same; that is, when we blame directly we will most often be motivated to blame in an affective manner. This is not to say that detached blame never occurs spontaneously, in fact, it might be the case that in many cases of third-party blame (discussed in a following sub-section) the blame is detached in nature. This might be the case as this type of blame might be a more low-stakes variety of blame, where the target of the blameworthy action is someone whose wellbeing one has no real stake in, for instance. While affective blame may be more common, and our impulse to blame is often affective in nature, detached blame does also occur and thus we must account for this type of blame as well.[8]

[8] Indeed, Pickard is not the only one to make such a distinction, however her framing of the distinction does give us a nice vocabulary to use here. Shoemaker & Vargas (2019) address a similar distinction. Specifically, they discuss dispassionate versus what one might call "passionate" blame. Shoemaker & Vargas' distinction is essentially the same as

It is important to note at the outset of this section, that not all accounts of blame are functional in orientation. For example, Scanlon (1998) treats blame as fundamentally a matter of moral appraisal: to blame someone is to judge that their action reflects a failure of judgment for which they are answerable. On such a view, blame need not essentially be communicative, forward-looking, or socially regulative; it may consist simply in holding a certain evaluative stance toward another's conduct. Similarly, Smith (2012) characterizes blame in terms of attributing a morally faulty quality of will. These accounts could quite nicely fit into the category of "detached blame," as the cognitive appraisals involved don't necessarily require the affective elements associated with other types of blame.

These views suggest that blame is centrally concerned with the assessment of agents rather than the regulation of social behavior. The norm-psychological account developed here does not deny this dimension of moral appraisal. Rather, it aims to explain why such practices of moral appraisal are so pervasive and stable: namely, because they are embedded within a broader system of norm enforcement, maintenance, and articulation. In this sense, my view is not offered as a competitor to appraisal-based accounts, but as a complementary explanation of why such forms of appraisal arise and persist.

2.2: Private Blame "in one's heart":

While blame can be either affective or detached, it can also differ along another dimension: that of expression. One type of blame that our leading theories often fail to account for is blame that goes unexpressed, or, private blame "in one's heart." As opposed to expressed blame, unexpressed blame would be constituted by those instances of blame that, for whatever reason, are not

Pickard's. That is, they focus on the ways in which blame can be detached and can exist without it's "characteristic sting."

expressed by the blamer. In these cases, it might be that the blamer simply cannot express their blame to the blamee (blaming the dead, for instance) and so the blame is kept completely internal. Or, in such cases, it could be that the blamer simply does not desire to express their blame to the blamee, or to a third party. Whatever the case may be, unexpressed blame is precisely that: blame which does not get expressed to anyone else. Now can unexpressed blame be affective or detached? Of course, it seems that it could be either. However, as previously discussed, it seems that many of our instances of blaming will be automatically affective in nature, and only when the blamer regulates their affective response do we get something like detached blame. As such, it might be the case that all internal or private blame also starts out as affective blame. Even if it is just for that first impulse of recognizing the phenomenon, it seems like there may be some affect attached in the first instance. However, it does seem that, with just a little bit of evaluation, that this type of blame could easily become detached as well. What might be the function of blame that goes entirely unexpressed? There is a question here to be answered here, but I will address this in a later section.

2.3: Direct Blame vs Third Party Blame:

One other dimension along which blame can differ is that of its directedness. Just as we can blame in a way that is expressed and also privately in our own minds, we can also blame individuals directly or we can blame indirectly to third parties. Blaming directly is what I have identified at the outset of the paper as paradigmatic blame. This occurs when one individual calls out another directly for some norm violation or breach. This type of direct blame is what we are discussing most often when we discuss blame. Third-party blame, on the other hand, occurs when an individual calls out some individual (the blamee), though not to that individual directly, but to some third parties who may or may not know or be socially involved with the blamee. Third-party

blame might end up looking quite similar to gossip, but with a slightly different tone. While the tone of gossip might be general interest, the tone of third-party blame is more instructive: alerting the relevant third parties to what the blamee has done and perhaps encouraging them not to engage with the blamee.

2.4: Self-Blame:

Another variety of blame that needs to be explained is self-blame. Self-blame is somewhat self-explanatory: it is blame that is directed at one's self. For example, one could imagine a chess player who fails to see a crucial move, loses the match, and blames themselves for it. Or quite differently, one could imagine someone who has a lunch date with their friend and gets so busy that they forget their engagement, standing up their friend. They might apologize to their friend, taking the blame: "I'm so sorry, I can't believe I forgot our lunch date, it's totally my fault."[9] It stands to reason that self-blame can be either affective or detached, and it can also be either private or expressed. One potential difference between self-blame and blaming others is that self-blame sometimes appears to target failures to live up to personal ideals rather than violations of shared social norms. However, even highly personal ideals often function as internalized normative standards. They structure self-evaluation and guide future conduct in ways analogous to socially shared norms. In this sense, cases of self-blame need not fall outside the norm-regulating framework; rather, they may reflect the enforcement of norms that have been internalized and taken seriously by the agent herself. In the two cases described above, the chess player and the forgetful friend, the highly personalized and internalized norms might be respectively something like the following. In the first case of the chess player, perhaps the norm that has been internalized is one of excellence: "I

[9] This may seem like a case of just regular blame, but I see no reason why self-blame can't be for something in which others were involved.

ought not have missed that move," or "I must do better next time." And in the case of the forgetful friend, it seems the norm is something like "I must not let my friends down." In instances of blaming others, we are often calling them out for some wrongdoing or norm violation. In cases of self-blame it seems that very often this is much the same. When we blame ourselves, we are marking out some norm violation that we have ourselves committed; one that we care about, and one that causes us to feel guilt, shame, or regret for having carried it out—incidentally or not.

So, our blame can be directed and overt, or third-party oriented. It can be affective or detached, and it can be expressed or private. While each of the leading theories of the function of blame can account for some of these varieties of blame, none of them can account for all of them. In the following section, I examine the current theories of blame and show where they all fall short, and the ways in which each has trouble accounting for some of these phenomena.

Section 3: Functional Theories of Blame[10]:

Theories of blame can be broadly divided along one important dimension. Some accounts, such as those developed by Scanlon (1998) and Smith (2012), treat blame as a form of moral appraisal tied to judgements about an agent's quality of will or failures of judgement. Other accounts emphasize the social roles that blame plays, treating it as communicative (McKenna 2012), as a form of protest (Hieronymi 2001), or as a costly social signal (Shoemaker & Vargas 2019).

While these approaches differ in emphasis, each captures an important aspect of our blaming practices. The challenge, however, is that no single account appears able to accommodate the full range of phenomena associated with blame. This motivates the search for a more unifying framework; one that can explain why blame exhibits both the evaluative features emphasized by

[10] Do note that many of the theorists discussed in this section would likely not describe their own theories as "functionalist" theories of blame or responsibility. What I take myself to be doing here, is identifying theorists who do in fact carve out various functions of blame, whether they acknowledge this themselves or not.

appraisal theorists and the social-functional features emphasized by more outward-looking accounts.

In this section I will discuss some functional theories of blame in the extant literature and where they fall short. Additionally, towards the end of this section I will discuss the functions of the abovementioned varieties of blame. I am focusing on functional theories of blame rather than, say emotional, or cognitive theories because, as I have shown, blame cannot merely be emotional or cognitive (it is often both). Indeed, blame cannot simply be an instance of anger or a mere judgement of wrongfulness. However, if we focus on functional theories of blame instead, we can start to see a very natural evolutionary picture arise. One thing to note at the outset of this section is that "function" can be taken in two ways here. First, we can discuss the current social functions that blame plays for us now. This is what the following functional theories of blame are referring to when they discuss blames' "function." However, we can also discuss a different type of function, specifically, the evolutionary function of the practice of blaming. While the following theories that I consider in this section all address the current social functions of blame, it could be the case that for our blaming practices, the current social functions and the evolutionary function are one and the same. It may be the case that the social-regulatory functions of blame are precisely what blame was evolutionarily selected for. Indeed, there is an interesting question here about the relationship between the current-social-function and the evolutionary function. My account does not claim that agents consciously aim at norm maintenance when they blame, nor that the justification of blame necessarily reduces to its evolutionary origins. Rather, the proposal is that blame is the kind of psychological and social mechanism that persists because it plays this stabilizing role. The explanatory level at which norm psychology operates is therefore complementary to, rather than competitive with, accounts that focus on the fittingness or interpersonal meaning of blame.

Perhaps it is the case that each of the following functional theories shows part of the adaptive nature of blame, but none gives us a full story. Each of the following theories does explain something about at least one of the social functions of blame, it is just the case that each theory also falls short. In the following I will show what some of the leading functional theories of blame illustrate, and also where each is lacking.

3.1: Blame as a call to behavioral change:

Hannah Pickard, in her paper "Responsibility Without Blame […]" gives a functional account of blame without calling it as much. In her paper, Pickard argues that affective blame is not clinically effective in the treatment of personality disorders and that detached blame is. Now, one thing we can take from this conclusion is that Pickard means for part of the function of detached blame to be behavioral change on the part of the wrongdoer. When clinicians blame wrongdoers in detached ways in the clinical context, this is part of effective treatment for these individuals with personality disorders. Because much of effective treatment for such individuals is grounded in behavioral changes, one can infer that Pickard views the function of detached blame as that of a request for behavioral change for the wrongdoer/patient. But this hardly seems like a complete picture of the function of blame. If one blames in a detached fashion, it could be the case that blame loses some aspects of its functions as a costly social signal. Indeed, it is the fact that detached blame is less punishing, that makes detached blame an effective part of treatment for individuals with personality disorders. Though affective blame might not work as a call to behavioral change in *this* clinical context, it seems that it may still serve this function in non-clinical contexts. As noted earlier, it seems that blame will have many functions and is a multifaceted mechanism that helps in the processes of group stabilization and group coherence. If

this is the case, then blame cannot *merely* be a mechanism of behavioral change. Indeed, it must be more than just that.

3.2: Blame as initiator of moral repair:

One other function of blame might be the expression of condemnation or disapproval of some action, which in turn should ideally lead to some form of moral repair. Michael McKenna (2012, 2013) argues that blame is *conversational*. For McKenna, blame functions to continue a conversation started by the blamee's wrongful action wherein one expresses the disapproval of that action. The ultimate function of such a conversation then, is to facilitate some change in the blamee; in expressing disapproval of the blamee's action the hope is for the blamee to recognize their wrongdoing and repair the situation (either through apology or some other means). McKenna's theory is a Strawsonian account, and thus grounds blame squarely in our morally reactive attitudes. Such reactive attitudes are not merely cognitive or conative, but are affective in nature. McKenna argues that certain morally reactive attitudes are directly implicated in blaming (McKenna 2013). Specifically, McKenna argues that the reactive attitudes involved in blaming are "best understood as a species of moral anger," and more specifically might look like resentment, and moral indignation (McKenna 2013). So, it seems that this functional theory cannot account for detached blame, nor for private blame or for third-party blame. If reactive attitudes are constitutive of blaming for McKenna, then it seems that this theory cannot account for the function of detached blame. Antony Duff has proposed a similar functional story about blame wherein expressions of blame function as attempts to communicate to the wrongdoer a moral understanding of their wrongdoing, i.e., to bring the wrongdoer to recognize their guilt and to repent for what they have done (Duff, 1986). But again, here it seems like these theories give us an incomplete picture of

blame's function. For, if blame can go unexpressed then it's function cannot merely be the communication of disapproval, or an initiator of moral repair.

3.3: Blame as a form of protest:

Pamela Hieronymi gives us another functional theory of blame, however, unfortunately, this account only addresses part of the picture of blame as well. Hieronymi (2001)—along with Matthew Talbert (2012), and Victoria McGreer (2013), who make similar proposals— argues that blame's function is a form of protest. The function of blame on this account is, most foundationally, to highlight whatever wrongdoing has been leveraged against the blamer and to mark it out as wrong. Specifically, Hieronymi offers that the reactive attitudes involved in blaming like resentment and anger (and the expression of these attitudes) serve as powerful forms of protest (Hieronymi 2001). Hieronymi articulates this theory of reactive attitudes and blaming within the framework of a theory of forgiveness. Ultimately, for Hieronymi it seems that how we blame is linked to how we ought to forgive, but forgiveness is a topic for another paper. Hieronymi describes three interrelated judgements that she believes are partially constitutive of blaming, together with affective responses:

> (1) The act in question was wrong; it was a serious offense worthy of moral attention. (2) the wrongdoer is a legitimate member of the moral community who can be expected not to do such things. As such she is someone to be held responsible and she is worth being upset by. (3) You, as the one wronged, ought not to be wronged. This sort of treatment stands as an offense to your person (Hieronymi, 2001).

Hieronymi argues that when these judgements are warranted, one's first response ought to be anger and resentment (Hieronymi, 2001). It seems for Hieronymi blaming is often undertaken in the hope of later forgiving; in order to forgive we first have to identify and call out the wrongdoing, e.g., in order to forgive we first have to blame the wrongdoer. According to Hieronymi our reactive attitudes serve as a powerful form of protest to wrongdoings that have been inflicted on us. These

reactive attitudes also seem to partially constitute blame for Hieronymi. One advantage of Hieronymi's view is that it captures the normative significance of blame in a way that purely instrumental accounts may miss. However, precisely because it ties blame so closely to these reactive attitudes and their expression, it struggles to accommodate certain cases where blame is either detached or entirely unexpressed.

It is here that we run into problems for this account. If blaming is necessarily linked to our reactive attitudes, then how can such a theory account for detached blame? If we *ought* to respond with anger and resentment, as Hieronymi claims, then how detached can we be in such responses (Hieronymi, 2001)? Additionally, there are many instances of blame where the blame itself is internal, i.e., the blame goes unexpressed. And it seems that it is paradigmatic of protest that it should be expressed. One commentator notes, "Do workers protest unfair labor conditions simply through their beliefs and attitudes? Or must they make such beliefs and attitudes known?" (Tognazzini & Coates, 2018). It seems like unexpressed protest would hardly count as protest at all, and yet there are many cases of unexpressed blame that still seem to count as blaming. So, it must be the case that blame goes beyond mere protest as its function as well.

3.4: Blame as a Costly Signal:

Shoemaker & Vargas (2019) propose a costly signal theory of blame. Specifically, they posit that blame is not a form of protest, or a call for behavioral change, but a costly signal sent by the blamer. Shoemaker & Vargas (2019) do a nice job of showing the way in which blame functions as a signal, but they fail to show how it functions in other important ways. Additionally, as mentioned earlier, this theory of blame will run into problems with our distinction between affective and detached blame as well. But, more on that later.

Shoemaker and Vargas argue that blame is, "[…] a costly response to norm violations defined most fundamentally not by any particular content—e.g., a mental state or activity—but by a *function*, namely, the signaling of the blamer's commitments, including a commitment to the enforcement of those commitments" (Shoemaker & Vargas 2019). For Shoemaker and Vargas this is most fundamentally the signal that blame sends but indeed, it is not the only one. We blame for many reasons and in many ways. Our blame takes many forms. We may directly blame the individual who has wronged us, and in these cases the costly signal is clearly directed at the blamed individual. This seems to be what Shoemaker and Vargas have in mind in this particular excerpt. However, we may also blame in other contexts where the costly signal is directed at a third party in our social group. In such cases of third party blame we may be signaling to others in our social group that we are not to be trifled with because we are willing to police certain types of norm violations, or alternatively that the person blamed is a morally bad actor who is willing to enact certain norm violations.

One aspect of Shoemaker and Vargas (2019) picture of blame is that it is a *costly* signal. Now what does this mean? Shoemaker and Vargas explain that, "In many instances, there seems to be no net payoff for blame. It often costs a lot, including emotional equanimity, time, energy, self-control, and self-governance. It can even cost the blamer dear friends and lovers." (Shoemaker & Vargas 2019). It is in this sense that blame is costly. It is a signal that involves the blamer taking on certain properties that are costly to themselves in some way. Shoemaker and Vargas (2019) maintain that, "When a costly signal becomes part of some stable system, it will be one in which it has observers, it is hard to fake (otherwise it would be too easily imitated), it delivers accurate information to the observers, and it benefits the signaler (Bird, Smith, & Bird, 2001)" (Shoemaker

& Vargas 2019). This is how Shoemaker & Vargas hold that blame functions in interpersonal scenarios. Shoemaker and Vargas (2019) consider a paradigm case:

> You've made a promise to help me move that, when the time comes, you simply blow off. The next time I see you, I angrily protest what you did, communicating to you a demand for acknowledgement and apology, and letting you know that I won't help you the next time you need help. In other words, I blame you. (Shoemaker & Vargas 2019).

Essentially, all of this activity in blaming signals multiple things to the blamee. First and foremost, such blaming signals to the blamee that it really does matter to the blamer that the blamee not break such promises, it additionally signals that the blamer is willing to enforce promises made to them. Shoemaker and Vargas note that, "This signal is hard to fake—hard enough that professional actors can fail to be fully convincing in capturing the involved attitudes, facial expressions, and bodily cues" (Shoemaker & Vargas 2019). It is important to note that part of what makes such a signal costly is also part of what makes it hard to fake. Shoemaker and Vargas (2019) discuss the costs embedded in such a signal:

> All sorts of unpleasant emotions have been stirred up in me. I have to invest time and energy in responding to you in this way, in carrying out or expressing the blame. I am also motivated to act in ways that involve risks, e.g., you, as the blamed agent, may respond in unexpected or nasty ways. I also risk the end or corrosion of our relationship. Blame occasionally costs one friends and loved ones (Shoemaker & Vargas 2019).

The fact that blame may be a signal that is difficult to fake tells us that it is not merely a costly signal, but it is also an honest and reliable signal. Shoemaker and Vargas (2019) discuss that emotions often play a key role in what makes costly signals so reliable: emotional reactions are often difficult to fake. So, it does seem like Shoemaker & Vargas (2019) accurately describe how blame functions as a signal. However, if blame is most fundamentally a signal of this type, then such a theory cannot explain blame that occurs privately "in one's heart." As we have established this is a common type of blaming that occurs, and so a good theory of blame needs to be able to account for it.

Furthermore, it seems like on the face of it, the costly signal theory of blame should not be able to readily account for detached blame. However, Shoemaker & Vargas do discuss what they identify as "dispassionate blame," albeit briefly. They address detached blame by providing an example:

> Consider a mother who has seen her adult son repeatedly cheat on his romantic partners. She may well blame him without rancor or any other passion: "You keep hurting others and yourself. You exhaust me. I don't know how much more of this I can take." This is an example of blame without reactive attitudes, relationship alteration, or protest […] (Shoemaker & Vargas 2019).

Some might not find this particular example compelling. Indeed, it is not entirely clear from this example whether this is truly detached blame or not. However, Shoemaker & Vargas argue that it is, and that the mother in this example is still fundamentally signaling. They even go so far as to state that, "the […] signal could actually be undermined if she were to signal her commitment to the violated norms with any reactive attitudes […]" (Shoemaker & Vargas, 2019). So, it seems that in this case the effectiveness of the signal is determined by the supposed absence of affect. While it seems that Shoemaker & Vargas can account for dispassionate / detached blame in this type of example, it seems that there may still be an open question of how costly such a signal is. If the signal that blame sends must be costly on this theory, then it is unclear whether detached blame is fully accounted for or not. Furthermore, if it is the case that it is blame's characteristic affective nature that makes the blame signal costly, then it seems that such a theory will have a bit more trouble accounting for detached blame than the authors purport to show here.

3.5: Functions of Private Blame:

Now, what might be the function of blame that is unexpressed? This type of blame certainly cannot function as a signal, a punishment, or a call for behavioral change, as these need expression for their proper function. If expressed blame is most often for the benefit of the

blamee (i.e. to signal to them that they should change or else people will not cooperate with them, or something like this), then it stands to reason that unexpressed blame might be just for the benefit of the *blamer*. Of course, unexpressed blame can play no function for the blamee, that they may be immediately aware of. If I blame you completely internally, then what this serves to accomplish is to motivate me, the blamer, in particular ways. Internal, or private blame functions as a motivation to disengage from potential future cooperation with the blamee. Now, I say that the blamee might not be aware of this function as they would have no way of knowing that they have been blamed and are being cooperatively disengaged with. In the sense that private blame can motivate the blamer to action, it seems that this variety of blame still consists in its function of behavioral regulation, it is just that in this type of case the blame is regulating the behavior of the blamer rather than the blamee. So, it seems that there is even a function for blame that goes completely unexpressed, and this is a motivation for the blamer to cooperatively disengage from the blamee.

3.6: Functions of Third-Party vs Direct Blame:

Now these two types of blame perhaps obviously serve different social functions. On the one hand we have direct blame which, as mentioned, seeks to change the behavior of the blamee, or calls out for some kind of relationship restoration, or apology, or something like this. That is, in cases of direct blame, the blamer seeks to alert and admonish the blamee of their norm violation, and seeks to encourage the blamee to change their behaviors. These types of cases of direct blame seem to be largely for the benefit of the blamee. It is important for the blamer to alert the blamee of their norm violation, and if they are not alerted to such they are liable to keep committing such norm violations. Should they continue to commit such norm violations they are likely to lose many

social benefits, like perhaps their place in the community as a whole (depending on the seriousness of their norm violations). This seems simple enough.

Now, on the other hand, cases of third-party blame might not be for the benefit of the blamee. As third-party blame is expressed to others in the community besides the blamee, the blamee may not even know that such third-party blame has occurred, and thus this cannot be for their benefit. Instead, third-party blame might be for the benefit of the third parties, that is, the other members of the community who may be negatively affected by the wrong actions of the blamee. For instance, third-party blame could function as a warning. The blamer blames the blamee to some third parties. In doing so the blamer has alerted these third parties that the blamee has enacted some such wrongdoing or norm breach, and is in a sense, warning them that this might occur again. And so, the function of third-party blame might be a warning to other community members that the blamee is capable of such bad actions and norm violations. Third-party blame might function as a signal to show others that they should be careful when cooperating with the blamee.

These are just some of the varied and many functional theories of blame. While each theory proposes some function and shows how this one purported function might work, each theory also struggles to account for some aspect of blaming. Communicative theories of blame cannot account for blame that goes unexpressed, theories that claim blame is a form of protest cannot account for detached blame, and the costly signal theory of blame also has trouble accounting for detached blame. It seems that each of these theories would have to do some "fancy dancing," as it were to account for such phenomena. But perhaps there is a theory that can account for blame's functions without such "fancy dancing."

The upshot of the discussion of these functional theories is not merely that blame serves many different functions, but that these functions appear to cluster in a systematic way. This suggests that we are not dealing with a loose collection of independent roles, but with a more unified underlying mechanism. In the following section, I argue that norm psychology provides the best explanatory framework for understanding this clustering, and for explaining why blame reliably exhibits these diverse but interconnected functions.

Section 4: Norm Psychology and Blame:

Humans are highly social beings. Unlike our primate ancestors and relatives, we often socialize and cooperate with individuals who do not belong to our immediate local group. In many cases we socialize with complete strangers who are previously unknown to us. In humans we also see large scale sociality and continuous social cooperation between non-related individuals. Additionally, we also find in human societies culture and cultural accumulation. With such widespread, large-scale cooperation and socialization there is bound to be conflict amongst individuals, and opportunities for free-riding. So, what are the stabilizing and cohering mechanisms that keep such large-scale socialization from devolving into chaos?

One such mechanism that keeps social groups together and functioning are norms. Norms of various kinds help guide individuals actions and help structure the world. Now, there is a further question here of what *maintains*, *manages, and articulates* the norms that guide our actions and shape our world. There are many mechanisms that maintain norms and blame certainly seems to be a candidate. So, what is interesting or important about the way in which blame maintains norms specifically? It might be the case that blame, in every form, is targeting norm violations. Blame targets norm violations when it is detached or affective, it does this whether it is expressed or unexpressed, and it does so when the blame is direct or third-party. Blame is punishing, it is a

signal, and it lends itself to reputational management, among other mechanisms. In the following sub-sections, I discuss how blame functions as a punishment, a signal, a tool of reputational management and more. The aim of my account here is not to reduce the normative significance of blame to its evolutionary origin or strictly to its evolutionary function alone. But it is instead to provide a framework for why such practices arise and persist. Questions about the justification or fittingness of blame remain distinct, even if they are informed by the kinds of mechanisms described below. My account can thus be understood as operating at a different explanatory level than many extant theories: rather than specifying the constitutive features of blame, it explains why practices with those features are so widespread and stable.

4.1: Norms and Psychology

Norms of various kinds are one stabilizing factor that help keep our large-scale socialization in order. From a very young age it seems that we can recognize and internalize norms of various kinds, and it is these norms which help the ultimate coherence of our social groups. Kelly and Setman (2021) state that, "Once a person adopts a norm, it functions both as a rule that guides behavior and as a standard against which behavior is evaluated" (Kelly & Setman 2021). Additionally, Kelly & Setman state that, "[…] individuals typically become motivated to enforce the norms they adopt, and so to participate in regulative practices such as punishment and the ascription of blame" (Kelly & Setman 2021). Such practices help to stabilize the overall social group's arrangements and the norms that structure them (Kelly & Setman 2021). In fact, it seems that our capacity to recognize, internalize, and enforce norms is so deeply ingrained that the supporting mechanisms will function well even when there is no direct benefit to anyone involved: "Importantly, these models/mechanisms (e.g. reputation, signaling, and punishment, among others) can stabilize any similarly costly behavior, strategy or practice, independent of whether it

delivers benefits to anyone; the costs of the action matter but the benefits are irrelevant for stability" (Chudek & Henrich 2011). Indeed, it seems that we are extremely sensitive to various norms, we internalize them easily, and their enforcement is natural to us:

> Young children show motivations to conform in front of peers, spontaneously infer the existence of social rules by observing them just once, react negatively to deviations by others to a rule they learned from just one observation, spontaneously sanction norm violators and selectively learn norms (that they later enforce) from older and more reliable informants (Chudek & Henrich 2011, references omitted).

As it turns out, once a norm has been internalized, it is also more difficult to break such a norm than it is to act in accordance with it: "Moreover, violating norms (i.e., breaking promises or inflicting harm) requires overriding automatic responses by brain regions responsible for cognitive control" (Chudek & Henrich 2011, references omitted). So not only are we extremely sensitive to norms and their enforcement, but their violation might actually be quite difficult for us to enact.

So, one can begin to see here how a theory of blame grounded in norm psychology might arise. Blame is one regulative practice that helps sustain the norms that guide our actions. Blame can be punishing, as we have seen given it's "characteristic sting,"; blame can regulate behavior and reputations given its generally communicative nature and given our propensity to blame to third-parties; and blame can (and often does) function as a signal. There seems to be a natural picture of blame here that both arises out of and can help accommodate our psychological theory of norms. In addition to being a signal, a punishment, and a means of communication, blame is a mechanism that helps *enforce* our norms of various kinds.

4.2: Punishments:

Now, let us turn to punishments. One outcome of a norm psychological account is the idea that individuals are quick to punish those who exhibit norm transgressions and breaches: "[…] prevalent norms and standards of conduct are collectively maintained by a community when its

members enforce them, punishing those who fail to follow the rules." (Kelly & Setman 2021). Indeed, we are so apt to punish those who violate various norms that we will do so even at a cost to ourselves (Kelly & Setman 2021). It seems very likely that blame might be one such means of punishment. When we blame others directly in dyadic blame scenarios (i.e., when we directly blame a wrongdoer) we are doing numerous things but one of the many things we are doing is punishing the wrongdoer for breaking some norm. Now what exactly makes blame punishing? On the one hand, it seems like it might be blame's generally affective nature that makes it so punishing. I believe this is part of the picture but not the full story. Of course, it will be the case that being blamed in an affective manner will likely feel punishing in nature. Expressions of anger, resentment, and disappointment, when directed at us, will generally feel unpleasant—and that in itself may be punishing. Not only this, but if blame is most often affective and is most often an expression of reactive attitudes like anger and resentment, then such attitudes will motivate actions that further constitute punishments. Anger motivates attacking or damaging the agent in some way or another, for instance, in blaming, perhaps by damaging the blamee's reputation. However, detached blame might also be punishing. If I blame you in a detached manner for some transgression just the mere threat of social devaluation might be enough for detached blame to still function as a punishment.

What about blame that takes place entirely internally on the part of the blamer? Can this variety of blame serve as a punishment? Perhaps, but only very indirectly. If blaming "in one's heart," is primarily for the benefit of the blamer, as we have previously established, then the way in which it is a punishment for the blamee is much less clear. As I have previously articulated, it might be the case that private blame functions as a motivation for the blamer to disengage from future endeavors with the blamee. If the blamee comes to realize that this is occurring, and that

certain individuals are no longer cooperating with them socially, then the social devaluation that results from private blame will perhaps be punishing.

Now, what about third-party blame? Indeed, our propensity to punish norm violations is so robust that we even do so when we have not been involved in them personally (Kelly & Setman 2021). It strikes me that third-party blame may be punishing in a similar way to private blame: very indirectly. Third-party blame may only serve such a function if the blamee comes to realize that certain individuals or groups are no longer socially cooperating with them.

4.3: Signaling

As mentioned in an earlier section blame can also function as a signal. When we blame, we are unintentionally signaling many different things to those around us. On the one hand in paradigmatic blame, we are signaling directly to the offender that we are not to be trifled with and we are willing to police such norm breaches as they have enacted. In such instances of blaming, we may also be signaling something along the lines of, "if you don't stop breaking this norm the group will no longer cooperate with you!" and as such blame might also be a request of sorts for behavioral change from the transgressor.

Signalling types of blame arise as a byproduct of our norm psychology plus combined with the motive to enhance one's status. Take cases of self-blame for instance. What might be signaled to others in cases of self-blame? Let's say I blame myself for some failure to live up to an ideal that I alone have set for myself, in blaming myself there is no connection to how others would or should respond to me, as it is upon no one but myself to blame me for my own failings to live up to my own ideals. However, as Shoemaker & Vargas (2019) note, self-blame often occurs in front of an audience, and thus it can still send a signal: "Think, for instance, of the quarterback who throws an interception and then pounds his helmet with his fists, the pro golfer who breaks her

club in frustration after a poor shot, or the tournament chess player who slumps in agony when he sees how he's just fallen into his opponents trap" (Shoemaker & Vargas, 2019). There are clear signals being sent here though they aren't being expressed verbally. Shoemaker & Vargas (2019) claim that the signal being sent in such cases is something like "I am committed to norms of excellence that I have failed to live up to" (Shoemaker & Vargas, 2019). Sending a signal of a commitment to personal excellence is a beneficial signal to send, according to Shoemaker & Vargas (2019). They argue that this is the case because people are attracted to those who are deeply committed to excellence (and to self-enforcing it) (Shoemaker & Vargas, 2019). It appears that even when blame is directed at the self it can still send a signal, though it will need an audience to do so.

Now, again here, one might wonder, what about private blame? What kind of signal does blame send when it is completely internal? It does seem like this potentially poses a problem as signals of norm commitment and enforcement appear to require an audience or a target to generate the benefits for the signaler (Shoemaker & Vargas, 2019). Shoemaker & Vargas (2019) again appeal to cases of self-blame to explain cases of internal blame, however it seems that here they miss the mark. Shoemaker & Vargas (2019), argue that many cases of self-blame lack an audience altogether, but this does not mean that there is no signal being sent in such cases, there is just no one around to receive it. Shoemaker & Vargas (2019) claim that sending such signals when no one is around to pick them up may make them particularly costly. In privately blaming, it may be that all of this affect has been stirred up in me, and yet I am signaling to no one. One can see that this is a particularly costly activity for the self-blamer, but one with little benefit. But, this doesn't quite give us the answer we are looking for here. It seems that Shoemaker & Vargas (2019) cannot

explain blame that takes place entirely internally, only in thought. This type of blame surely cannot be a signal of anything.

4.4: Behavioral Change and Norm Articulation:

In addition to punishment and signaling, blame can also function in ways emphasized by forward-looking theorists. On such views, practices of responsibility, including blame, contribute to shaping and scaffolding moral agency over time. By holding individuals answerable for their actions, we participate in the cultivation of more reliable norm-compliers and more reflective agents. Not only this, but we often, at the very same time, use blame (and other such practices) to articulate new norms, and enforce them simultaneously. Kelly & Westra (2026) describe a case that both constitutes blaming, norm articulation, and norm enforcement all at once. In their talk, "The Psychology of Articulated Norms," they describe a case of a professor, who misgenders a student (the student uses they/them pronouns, and the professor mistakenly refers to them as "she,"). Another student raises her hand, and corrects the professor, saying "they." Kelly & Westra (2026) use this case as an example of norm articulation, and a case where the professor's old, internalized norms around gender and pronouns are bumping up against the changing tide of the new norms around pronouns that the students want the professor to internalize. Not only is the correcting-student articulating a new norm, and enforcing it at the same time, but by my estimation she is also *blaming* the professor by issuing a signal and a social sanction.

A norm psychological framework can accommodate these insights. If blame is one mechanism through which norms are stabilized, articulated, and reinforced, then its role in shaping agency is not an alternative to norm regulation but one of its central expressions. Blame contributes not merely to the immediate correction of behavior, but to the longer-term maintenance and articulation of shared normative expectations within a community.

4.5: Reputational Management:

In addition to the familiar functions of blame discussed and explained above, a norm-psychological account enables us to see yet further functions of blame. One of which is reputational management. On such an account, blame will have this function for both the blamer and the blamee, though of course the function will look different in each respective case. Let's take a case of paradigmatic blame as I have been calling it. Recall the example from the very beginning of the paper: I promise you that I will help you move, and then I leave you high and dry when the time comes. You, rightly, blame me for this. But let's suppose further that you blame me in front of an audience of our peers. In blaming me in this way you are doing all of the things just discussed: you are signaling your normative commitments both to me and to the present audience, you are punishing me directly for violating a norm, and you are implicitly (or perhaps explicitly) asking me to change my behaviors or face further consequences. These actions on behalf of the blamer (you, in this example) will likely cause certain reactions in the audience or present third parties, for example, distrust that I (the blamee) will hold up my end of the deal in such promises and commitments. These reactions will affect how the third parties come to view the blamee, e.g., such reactions will probably have an affect on the blamee's reputational status! Not only this but blaming in this way may also affect how the present third parties view the *blamer.* Witnessing one agent blame another will likely modulate the blamer's reputation as well as the blamee's. Blaming shows not only that the blamer is committed to the norms that have been broken, but also that they are willing to police such breaches. This might give them a particular reputational status— that of one who is willing to enforce norms. And thus, they may stand out in their community accordingly. However, these needn't be the only responses that third parties might have to witnessing the blaming of another agent. In certain cases, it seems plausible that witnessing someone being blamed for some norm

violation could evoke sympathy for the blame. For instance, if the third parties present have also been blamed for some similar wrongdoing, they might change their personal reputational status of the blamee, but also feel sympathy for them. Of course, the reputational management aspect for the blamee might also be mitigated by the blamee's response to being blamed. If the blamee is sufficiently repentant and contrite, perhaps if they apologize sincerely, then their reputation may be less damaged than it might be otherwise.

Section 5: Conclusion:

So, blame is primarily important because it is one mechanism among many that helps maintain and manage our norms of various kinds. Blame assists us in enforcing norms, in punishing norm violations, and in signaling our commitments to certain norms. As such, is an important stabilizing and organizing feature of our social ecology. By this point it should be clear that most of the leading theories of blame have some trouble accounting for some of its functions or aspects in some way or another. Accounts that suppose blame is a communicative device cannot account for unexpressed blame. The protest account of blame also has trouble with this same task. The costly signal theory of blame has trouble accounting for private blame, and it cannot so easily account for detached blame, as the affective and expressive components of blame are what make it a costly signal. But again, each extant theory of blame also gets something correct about blame. Accounts that claim that blame is a communicative device are correct that expressions of blame are trying to communicate something to the wrongdoer, specifically, they are trying to show the wrongdoer that their action was in violation of some norm and will not be tolerated. The protest account of blame is right to point out that our blame and the reactive attitudes associated with it is a powerful way to call out the wrongdoing that has been done to us or to others. The costly signal theory of blame is also correct that blame functions not just as a social signal, but as a costly one!

Each account does seem to get something right, it is just that none of them can fully account for all of blame's disparate and different functions. When we look to norm psychology, however, we find that we can indeed account for blame's many different functions and facets. On the norm psychological view, existing theories of blame can be understood as idenfiying different functional aspects of a single practice, rather than as offering mutually exclusive accounts. As hyper social beings, we need stabilizing and organizing mechanisms to facilitate prosociality. Blame is one such organizing and stabilizing mechanism.

Chapter 4: Acceptable Apologies

Section 1: Introduction:

In ordinary experience we often make a distinction between acceptance of an apology and forgiveness of a wrongdoer. I take it as uncontroversial that one might reasonably respond to an apology by saying, "I accept your apology, but I don't forgive you." There is a puzzle here that needs attention: in the extant literature it is often assumed that forgiveness comes part and parcel with the acceptance of a sincere apology, however descriptively this is not always the case. In the following I will carve out such a distinction. That is, I will argue that there is a morally significant difference between the acceptance of an apology and the forgiveness of a wrongdoer. These two acts needn't be connected, though they sometimes are. One can forgive when no apology has taken place, and one can rationally accept an apology without forgiving. Contrary to some accounts of apologizing, no matter how good an apology may be one can never incur an obligation to forgive. *Forgiveness,* then, is supererogatory. However, I suspect that there may be cases where *accepting* a good apology is the morally right thing to do. If this is true, there is a further question of what exactly makes an apology worth accepting. What features make good apologies good, or rather, acceptable? And further still, when is an apology complete? I will argue that there are certain conditions that make an apology acceptable, and that should such conditions be met by the wrongdoer, then it might be the morally correct thing for the victim to accept such an apology. I will further contend that such acceptance begets completeness in the process of apologizing.

Most extant theories of guilt characterize apologies as one type of restorative action that the wrongdoer can take to make things right with the victim. On the other hand, in theories of resentment, forgiveness is often one way to the alleviation of such a negative emotion for the

victim. In apologizing a good moral agent should not apologize with the hopes of alleviating her feelings of guilt, but she should instead do so as part of a sincere attempt at restoration. The wrongdoer cannot simply apologize to discharge her feelings of guilt, she must also ensure that she has no further reason to feel guilty. Likewise, if a victim forgives simply to rid herself of her feelings of resentment, then she is—to use Pamela Heironymi's terminology—compromising in her forgiveness.[11] Any good account of apologizing and forgiving must make it such that all parties involved are acting sincerely and for the right kinds of reasons. Such an account cannot only articulate what the wrongdoer can do in the process of apologizing well— it must go further than this. A good account of apologizing must also make it clear that the victim cannot incur a duty to forgive or to forswear resentment. A victim of a wrongdoing cannot be in a position to incur moral blame for failing to forswear her resentment. However, such an account must allow for the victim to accept the wrongdoer's sincere apology and begin the process of restoration, while still rationally feeling resentment towards her wrongdoer.[12] The account I will articulate makes space for the victim to accept and complete the wrongdoer's sincere apology, and also allows for the victim to uphold her uncompromising forgiveness, without adopting any undue obligations.

The main purpose of my discussion here is to articulate an important distinction between nodes on a continuum of moral repair. I intend to make my argument here in two larger overarching parts. In this first part of the paper, I argue that forgiveness and acceptance occupy different grades of supererogation. To do this, I appeal to Pamela Hieronymi's account of "uncompromising forgiveness." I also argue that my view can account for the view of forgiveness Hieronymi sets

[11] Hieronymi 2001

[12] My account can *allow* for such cases, but there can certainly also be cases where acceptance and forgiveness do not come apart in the way I am articulating. Of course, there can be cases when a victim accepts an apology and forgives at the same time. My account simply creates more room for the cases we often face in ordinary life where this does not occur.

hers against, or what I call the "orthodox view." The orthodox view is that forgiveness entails—or simply is—the forswearing of resentment. I argue that forgiveness involves two distinct mental processes: one that is voluntary, and one that is not. The voluntaristic side of forgiveness looks much like Hieronymi's account—a cognitive reappraisal of one's moral judgements about the wrongdoer or the ways in which one has been wronged. The involuntary side of forgiveness then is the affective shift that occurs either with time, or as a result of the wrongdoer holding up her end of the promise she makes in apologizing. However, while a psychological change in attitude from resentment to a positive or neutrally valanced mental state might constitute some form of forgiveness, that which only occurs attitudinally is not within the realm of supererogation. The communication to the wrongdoer that such a change in attitude has taken place—i.e., the actual speech act of forgiving—is what we should instead regard as supererogatory. I hold that one can accept an apology and that one needn't, or in many cases one simply cannot, forswear resentment at that time. Acceptance, as opposed to forgiveness, might look like the willingness to *trust* the wrongdoer in her promissory apology.

In the second part of this paper, I will respond to Adrienne Martin's Strawsonian account of the good-making features of apologies. I will also argue that many interpersonal apologies are, in fact, much like promises. Just as promises create an obligation on the part of the promisor, so do apologies create a similar obligation for the wrongdoer. Explicitly, if interpersonal apologies are in effect promises of better behavior, then in apologizing the wrongdoer generates an obligation to hold up her end of the deal. I will call these types of apologies, "promissory apologies," and my discussion here will be limited to these. The more interesting point, I hold, is that in cases of sincere promissory apologies, it may be the morally right thing for the victim to *accept* an apology that

meets the revisionary good-making conditions I will ultimately propose. However, these conditions need argumentative support.

The incorporation of acceptance as an intermediary point between the rejection of an apology and the forgiveness of the wrongdoer ultimately allows not only for (1) the completion of the wrongdoer's apology wherein the process of restoration can begin, but also for (2) the victim to have more cognitive space to rationally process her anger and resentment, and to do so without an undue obligation to *forgive* the wrongdoer before the proposed restorative process takes place.

Section 2: Acceptance and Forgiveness:

The distinction I want to defend can be captured in a familiar scenario. When someone offers us an apology, we might reasonably respond by saying, "I accept your apology, but I don't forgive you." This response suggests that acceptance and forgiveness are distinct moral acts. Yet, most philosophical accounts treat them as equivalent or assume that accepting a sincere apology necessarily involves forgiving the wrongdoer. I argue that this conflation obscures an important moral distinction. Acceptance and forgiveness occupy different grades of supererogation, and recognizing this difference has significant implications for how we understand both the obligations of wrongdoers and the autonomy of victims. When faced with a sincere apology that promises better behavior, it might be—to borrow Elizabeth Harman's terminology— a "morally permissible moral mistake," to refuse to accept.[13] However, the same cannot be said of forgiveness, which I argue occupies a different grade of supererogation. What makes this distinction philosophically important is that it allows us to complete the moral work of apologizing without requiring victims to forswear their resentment prematurely (and, in fact, as I will argue, they may not be able to do

[13] Harman (2016)

so in some cases). Acceptance can serve as an intermediary response that acknowledges the wrongdoer's sincere effort at repair while preserving the victim's emotional and moral autonomy.

I hold that forgiveness involves two distinct mental processes: one that is voluntary, and one that is not. The voluntaristic side of forgiveness looks much like the orthodox account—a cognitive reappraisal of one's moral judgements about the wrongdoer or the ways in which one has been wronged. The involuntary side of forgiveness, then, is the affective shift that occurs either with time, or as a result of the wrongdoer holding up her end of the promise she makes in apologizing. However, while a psychological change in attitude from resentment to a positive or neutrally valanced mental state might constitute some form of forgiveness, that which only occurs in the mind of the victim cannot be within the realm of supererogation. It is instead, the *communication* to the wrongdoer that such a change in affect has taken place—i.e., the actual speech act of forgiving—that we should regard as supererogatory. I hold that one can accept an apology and that one needn't, or in many cases one *simply cannot,* forswear resentment at that time. The acceptance of a sincere apology allows the victim to rationally feel resentment towards the wrongdoer, or the wrong action for as long as necessary.

This is contrary to most of our orthodox accounts of forgiveness. Most accounts of forgiveness do not make this distinction at all. Philosophical accounts of forgiveness usually assume that if an apology is sincere and is accepted by the victim, then the victim also forswears her resentment and forgives the wrongdoer. In the following I will briefly sketch a few views of forgiveness, and I will then illustrate how acceptance functions relative to such views.

2.1: The Orthodox View:

For many theorists in the forgiveness literature, forgiveness simply is the forswearing of resentment.[14] The claim that forgiveness is the forswearing of resentment is not a new idea. The idea that forgiveness is the forswearing of resentment purports that if we are good moral agents, then we should strive to be virtuous. On the orthodox view, forgiveness is a virtue towards which one should strive; it is an *intentional* forswearing of resentment. It is a cognitive reappraisal and change in judgement in light of a sincere apology. It is something that we should strive to do and something that we must do intentionally: we should want to be forgiving. This is, more or less, the "orthodox" view of forgiveness.[15] However, it should be clear that the orthodox view only encompasses one of the two components of forgiveness that I have identified: the voluntaristic component. The orthodox view cannot account for the involuntary shift in affect that the victim feels, which may or may not accompany the cognitive reappraisal of her moral judgements about the harm or wrongdoing that has occurred. Whether or not these two components of forgiveness are co-extensive will often depend on the individual victim's ability to successfully regulate her emotions in tandem with her moral judgements. If acceptance is included in the moral framework, then forgiveness can occupy a higher grade of supererogation, and the completion of a good apology can still be possible.

2.2: Accommodating an Uncompromising Forgiveness:

Another view of forgiveness that we now must briefly turn to is Hieronymi's view posited in her paper, "Articulating an Uncompromising Forgiveness." Hieronymi endorses an account of resentment that takes the object of resentment to be the insult or claim launched at the victim as a

[14] Moran (2013), Hieronymi (2001), Calhoun (1992), and others.

result of the wronging done to her. The main part of Hieronymi's argument that I want to address here, are the three judgements that the victim should not compromise in forgiving. Hieronymi argues that in uncompromising forgiveness, the victim must be able to uphold the following three judgements:

> (1) the act in question was wrong; it was a serious offense, worthy of moral attention.
> (2) the wrongdoer is a legitimate member of the moral community who can be expected not to do such things.
> (3) you, as the one wronged, ought not to be wronged. This sort of treatment stands as an offense to you.[16]

I do think that this account makes sense. For Hieronymi the object of the victim's resentment is the insult generated by such an offense. And in apologizing the wrongdoer attempts to erase said insult. Ultimately though, while Hieronymi's account is quite convincing, it equivocates on the acceptance of an apology and the forgiveness of wrongdoers. She states, "In accepting the apology, the offended in some way ratifies, or makes real, the offender's change in heart," and later in the same paragraph, "If all goes well, the joint action of requesting and granting forgiveness will leave the original meaning of the event in the past."[17] While Hieronymi does use the terminology "accepting an apology," she does not state that this is in any way different than granting forgiveness, and additionally seems to use the two phrases interchangeably in her paper. So, it seems that we find here, an instance where this distinction is hinted at, but not clearly defined.

Further, in articulating her view of forgiveness Hieronymi states, "I will only examine the sort of forgiveness we have been considering: that in which an apology brings about a change in view or revision in judgement that allows one to forgo resentment."[18] However, Hieronymi also argues that she does agree with Cheshire Calhoun's account of "elective forgiveness," which is to

[16] Hieronymi (2001) pp.530
[17] Ibid. pp.550
[18] Ibid. pp.545

say that forgiveness cannot be obligated or demanded of others—it must be something we freely give. In this sense Hieronymi endorses that forgiveness is supererogatory, but she also notes that there may be cases where refusing to forgive might not be the morally right thing to do either. Hieronymi states that, "[…] there is certainly a sense in which being "unforgiving," rightly draws blame, in which we do owe one another forgiveness."[19] This is precisely why the incorporation of acceptance as an intermediary node on a continuum of moral repair is necessary. In such cases where it, *prima facie* seems that being unforgiving might incur blame, what might more accurately be occurring is a refusal to *accept* a well-made apology. However, I want to note that rather than being deemed blameworthy for refusing to *accept* a good apology, or for refusing to forgive, we might rather call these instances different grades of failures of supererogation. To turn to the analogy of promissory obligation briefly, just as some promises hold more moral weight than others (e.g., a promise to meet for lunch versus a marriage vow), supererogatory acts may also differ in their requisite grade of supererogation. To leverage Elizabeth Harman's terminology, I propose that if the conditions that I will develop in §3 are met by a wrongdoer in apologizing, then it might be a "mere moral mistake," for the victim not to *accept*. However, it is less clear to me that a victim's refusal to forgive could be carved out in the same fashion. In alignment with Calhoun's view, I posit that forgiveness (or at least the communication of forgiveness to the wrongdoer) is always supererogatory, and that it is not even a morally permissible moral mistake to refuse to forgive if the victim should choose to do so. For, again, there may be cases in which it is impossible for the victim to fulfil both the voluntary and the involuntary aspects of forgiveness at the same time. If part of forgiving is an affective shift, then it must be the case that, at least sometimes, we cannot all forgive willingly. Even if we can change our minds at a particular time,

[19] Ibid. pp.552

we cannot necessarily change our affective state. However, we can always choose to accept an apology, should it be sincere. It is by way of the inclusion of acceptance into the moral schema that we can better accommodate Hieronymi's uncompromising forgiveness.

Section 3: Accounts of Apologizing:

In taking up the question of what successful apologizing entails, we will have to pin down exactly what features make a good apology good and why. I contend that these conditions will be slightly different given the different social facts about each interpersonal situation requiring an apology. While an apology between individuals might aim toward restoration of their relationship, a good public apology might aim towards something larger, perhaps something in the arena of restorative justice or reparations, depending on the wrong that has been done. For the purposes of my project here, I will only discuss interpersonal apologies in cases where some actual wrongdoing has taken place.

3.1: Martin's Account:

I will posit that some conditions for a successful apology can be applied generally across most cases of interpersonal conflict or wrongdoings. These conditions do, in some sense, correspond to the "normative expectations" that Adrienne Martin explains will be relevant in cases of resentment, but more needs to be said here. In her paper, "Owning up and Lowering Down: The Power of Apology" Martin gives a Strawsonian account apologizing, resentment, and forgiveness. Martin posits her Strawsonian account in contrast to the insult account of apologizing and forgiveness. Attributed to Pamela Hieronymi and Jeffrey Murphy respectively,[20] the insult account

[20] Hieronymi and Murphy are not the only proponents of this view. But they are who Martin is specifically arguing against. However, the way that they each cash out this account is slightly different. I will focus here on Hieronymi's account of forgiveness and resentment, and Martin's reading of such.

focuses on the message or claim that the wrongdoing sends to the victim: i.e., an insult— that you as the victim deserve such bad treatment. Indeed, some theorists insist that the guilt of the wrongdoer, and the resentment of the victim take the same object, but what that object is stated to be is more controversial.[21] Hieronymi, as a stated proponent of the so called insult account takes the shared object of the wrongdoer's guilt and the victim's resentment to be the claim that the wronging creates.[22] Martin however, does not posit a shared object of the wrongdoers guilt and the victim's resentment. It seems right to me to say that the wrongdoer's guilt and the victim's resentment should share the same object. However, whether this object is the insult or claim that the wronging creates, or the actual wrong act itself is inconsequential to my view here. According to the insult account of resentment, a sincere apology acts to effectively erase the insult of the claim the wronging imposes on the victim—this happens through the process of the victim's reappraisal of her feelings of resentment, resulting in a change of her judgement about the situation or about the wrongdoer.[23] But as Hieronymi argues, this must happen for the right kinds of reasons.

In her paper, Martin poses the question of how an apology can give a victim reason to maintain a relationship with the wrongdoer, while also rationally refusing to forgive.[24] Additionally, she contends that the insult account cannot accommodate this descriptively accurate outcome. Ultimately, Martin draws two conclusions: (1) that sincere apologies succeed in providing the victim a reason (or, further, an obligation) to forgive, and (2) that some apologies provide reason for the victim to maintain her relationship with the wrongdoer while at the same

[21] It could also be interesting to examine cases where the wrongdoer's guilt and the victim's resentment do not take the same object. For instance, cases where perhaps the wrongdoer apologizes for causing some emotional harm to the victim, but does not feel that her actions were wrong, and thus does not feel guilt. But perhaps this is better saved for another time.
[22] Hieronymi (2001)
[23] Ibid.
[24] Martin (2010)

time refusing to forgive.[25] In making my own positive argument, I will show that as forgiveness is supererogatory (1) can perhaps only partially obtain; we can have reasons to accept a very good apology but we cannot have an obligation to forgive. And I will give an amended account of apologizing and *acceptance* that can more adequately account for (2). However, acceptance of an apology does not necessarily entail that the victim maintains a relationship with the wrongdoer, but my view can accommodate this outcome.

Martin contends that the wrongdoer must do three things for an apology to meet her good-making conditions. Martin's conditions are as follows:

> (a) own one's actions,
> (b) express sincere regret, and
> (c) express an attempt to repair the wrongdoing or the relationship.[26]

Martin also explains that there must necessarily be a performative aspect of a successful apology. I think this is essentially right, but that even the minimal conditions for a successful interpersonal apology need to go a bit further than what she requires. Martin also argues that apologies that meet these conditions are powerfully reason-giving.[27] Specifically, she argues, in coherence with (1) stated above, that "being the victim of a wrongdoing rarely if ever puts one fully beyond the call of reasons generated by the wrongdoer's attempts to rectify the situation."[28] I take this to mean that if an apology meets the conditions that Martin proposes, then such an apology should have so much reason-giving force that the victim is then *required* to forgive the wrongdoer.[29] Martin characterizes apologies of this type as marking out the moral superiority of the victim over the

---

[25] Ibid. pp.537
[26] Ibid.
[27] Ibid. pp.534
[28] Ibid. pp.547
[29] While Martin does not explicitly state that victims are never put beyond the call of *duty* in forgiving their wrongdoers, I am not sure how else to make sense of her claim (excerpted in the previous sentence). Given the context and her other arguments, it is safe to think that Martin might be equivocating "beyond the call of *reasons*," with, "beyond the call of *duty*." So, I take it that what Martin means to argue here, is that sincere apologies often do generate an obligation for the victim to forgive the wrongdoer.

wrongdoer. As the morally superior agent, the victim is required to forgive the wrongdoer out of some duty to help those who are morally inferior or morally worse off. Further still, Martin contends that if her definition of a good apology is correct, then such an apology can "provide good reason to forswear resentment, therefore making it unreasonable or irrational to refuse to forgive."[30] It strikes me that her account is in need of revision. A good account of apologizing cannot create an obligation of forgiveness on the part of the victim, nor can it make it irrational for a victim to refuse to forgive. However, it could be possible that there are better conditions for a successful apology that make it such that the victim's refusal to *accept* turns out to be a "morally permissible moral mistake."[31] Such conditions should, in turn, also make it the case that by refusing to compromise in her forgiveness while—at the same time—accepting the wrongdoer's apology, the victim is not deemed irrational, or morally blameworthy.

3.2: Promissory Apologies:

Let's turn now to the connection between apologizing and promising. In her account of apologizing, Martin also mentions the relation to promises but, again, all too briefly. Martin argues that an apology is only sometimes like a promise, and that even in promise-like apologies there is no "pseudo-contract," that the wrongdoer and victim enter into in the process of restoration. For now, it is simply important to note that Martin argues both that interpersonal apologies do not generally function like promises, and yet perhaps opposingly, that good apologies can generate an obligation for the victim to forgive the wrongdoer.

Contra Martin then, I claim that insofar as many interpersonal apologies *are* in effect, a promise of better behavior, that the wrongdoer and the victim *do* enter into some sort of pseudo-

[30] Ibid. pp.547
[31] Harman (2016)

contract as a result. For example, take a rather benign case of promissory obligation: if I promise to you that I will meet you for lunch tomorrow, I generate an obligation for myself to fulfill the conditions of this promise. This is to mean that you, the promisee, may be rightly disappointed, upset, or resentful if I do not hold up my end of the promise, e.g., if I do not show up for our lunch. Now, of course it can be the case that some promises are more or less morally weighty than others. The promise to go to lunch with a friend carries less moral weight than a promise like a marriage vow, or a promise to return your friend's car unscathed after borrowing it. But in all cases an obligation is generated.

In her paper "The Problem with Sexual Promises," Hallie Liberto argues that positive and negative sexual promises generate not one, but two obligations: one on the part of the promisor (as is the case for most promises), and one on the part of the promisee. The obligation incurred by the promisee, on Liberto's view, is an obligation to *refuse* to accept such a promise, or if such a promise is accepted, then the promisee has a further obligation to release the promisor from her promise should such an occasion arise.[32] I hold that interpersonal apologies function in a somewhat similar way. Like promising, in sincerely apologizing to you—assuming my apology meets the conditions that will make it a good apology— I create an obligation for myself as the wrongdoer to hold up my end of the deal (i.e., to actually follow through on the things I say I will do in my apology). Not only this, but if my apology meets the conditions that I will propose, then it might be the morally right thing for you as the victim to *accept* my apology. Contrary to Martin, however, I do not think that even the best of apologies can create a requirement for the victim to forgive.

I want to make it clear that in refusing to accept such an apology a victim cannot be found blameworthy, but it could still be the morally right thing in these cases for a victim to *accept*. What

[32] Liberto (2017)

should we take this seemingly contradictory claim to mean? To borrow Elizabeth Harman's terminology, it would be a "morally permissible moral mistake," for the victim to refuse such an apology. Harman argues that acts which occupy the category of "morally permissible moral mistakes," are those that are supererogatory— and thus are not obligations— but they are also sometimes the actions that are the morally right choices. Given all our reasons for action that bear on a situation, in these cases choosing not to ϕ would constitute a "moral mistake."[33] The refusal of accepting sincere promissory apologies, then, occupies this category of moral action— it would be morally wrong for the victim to refuse to accept. However, no one would find her morally *blameworthy* in such a refusal. Contrary to Martin, if forgiveness is supererogatory, then even the best of apologies cannot be so forcefully reason-giving that they require a victim to *forgive*. I contend then, that acceptance and forgiveness occupy different grades of supererogation.

Section 4: Success Conditions for a Good Apology:

Now, what exactly makes a good apology good? As discussed in §2, Martin's proposed three conditions for a good apology are that the wrongdoer should: (a) own one's actions, (b) express sincere regret, and (c) express an attempt to repair the wrongdoing or the relationship.[34] Again, an account of successful apologizing needs to go further in the conditions that it articulates.

The conditions I propose are by no means exhaustive of what makes a good apology, but they are merely the minimal conditions that might satisfy this concern. Do note that conditions 1-3 may be necessary for a successful apology. However, condition 4, I believe, will be relevant in most but perhaps not all successful apologies. And likewise, the nature of condition 5 makes it such that it will vary dependent on the social facts and moral context in which the transgression

[33] Martin (2010)
[34] Martin (2010) pp.537

and apology occur. All of this being said, I will list my proposed conditions for an acceptable apology, and then in the following subsections I will elaborate on each. My proposed conditions are as follows: performativity, an attitude of unambiguous guilt, sincerity, allocution, and an agent-relative condition that allows the victim to impose agent-relative demands on the success of the apology.

It is important to pause here and note that the same types of circumstances that would override the moral validity of, say, a promise, will also override the moral validity of a good apology. The conditions under which it might be the morally right thing to *accept* an apology cannot obtain in cases of active abuse, victim shaming, coercion, threats, blackmail, or anything like this. To give an extreme example, if a wrongdoer holds a gun to the victim's head and gives an excellent apology that meets all of the abovementioned conditions, but this apology occurs under such circumstances and ends with, "now, accept my apology or I'll shoot!" then this obviously cannot qualify as a good apology.

The conditions under which an interpersonal apology can succeed must be just stringent enough. If such conditions are too broad, then we run the risk of individuals over-apologizing or apologizing for the wrong kinds of reasons. A good theory of apologizing needs to account for these things. Likewise, if the success conditions for a good apology are too strict, then it could make it the case that no one could successfully apologize at all, and so this also cannot be the aim. My goal in postulating these conditions is again, not to create an exhaustive list of conditions for good apologizing, but just to give some minimal conditions that might render an apology acceptable. In the following sub-sections, I will address each condition in further detail.

4.1: The Performativity Condition:

For an apology to be successful, there must be a performative component to the apology rendered in public (i.e., generally in the form of a speech act). The idea that there must be some publicly rendered component of successful and acceptable apologies seems *prima facie* intuitive. However, there could be cases where this does not happen and in postulating such a condition it is important to address this. Additionally, there are certainly many cases of *forgiveness* that do not involve a performative speech act—i.e., those wherein the victim never encounters the wrongdoer again and yet chooses to forgive in her own mind. There is a question here as to whether this type of forgiveness can still be said to be supererogatory, but I will say more on this later.

If we appeal to some of the literature on permissive consent the reason why this might not always be an intuitive condition becomes clear. In the extant literature in the philosophy of consent, there are two overarching and competing theories of how consent might be rendered between individuals. Of course, this applies most often to cases of sexual consent, but it can apply to other types of consent as well. The two accounts are the performative account of consent, and the attitudinal account. In considering why there must be a performativity condition for good apologies, I will perhaps surprisingly, examine the account that might *preclude* such a condition; namely, the attitudinal account. Proponents of the attitudinal account of rendering consent hold that consent is an act that occurs in the mind of the right holder, and that one needn't publicly render one's consent for it to be morally valid.[35] Theorists who uphold this view claim that it advances the agent's autonomy for consent to be a private or internal event. Attitudinalists about consent argue that this account is autonomy enhancing because the if agent consents privately, then she cannot be influenced by any external coercion or persuasion. So, one could imagine

[35] Alexander (2009); McGregor (2005).

analogous cases of apologizing. Perhaps there could be a case wherein the wrongdoer feels so much shame, rather than guilt in response to the wrong she has done that she simply isolates herself, but she *feels* very sorry for the harm she has caused. It goes without saying that this cannot meet the desideratum of what a good apology calls for. Without the performative aspect of apologizing, it is unclear that a bare "apologetic" attitude could be considered an apology at all. However, because such cases might be possible, or because a reader might object to a performativity condition on these grounds, the reasoning for such a condition still must be stated. If I am correct in arguing that apologies are in effect, promises of better behavior, then it is clear that a wrongdoer cannot sincerely apologize without extending some sort of performative speech act to the victim. However, since part of my claim is that forgiveness is not required in rendering an apology complete, there can certainly be cases of *forgiveness* that occur only within the mind of the victim.

4.2: The Attitudinal/Affective Condition:

A good apology cannot *merely* be a performative measure on the part of the wrongdoer, there must be both a performative component *and* an affective or attitudinal requirement for an apology to be acceptable. As is the case with many of our best theories of moral actions, I cannot address such questions without also articulating the relevant emotions, motivations, and reasons involved in the processes of apologizing and forgiving. Specifically, the extant literature on apologies associates sincere guilt, shame, and remorse with the wrongdoer, and likewise the current literature on forgiveness attributes resentment, anger, and distrust to the victim. The attitudinal, or affective condition helps accommodate such a discussion. Apologizing and forgiving are necessarily intertwined with the motivations, judgements, and emotions of both the wrongdoer and the victim. On the part of the wrongdoer, we think sincere unambiguous guilt or remorse are

the types of mental states that will make an apology acceptable. There is a further question of how to gauge the sincerity of the wrongdoer's *guilt*. However, this question is better answered by the following allocution requirement.[36] The extant psychological and philosophical literature on guilt, shame, and apologizing purport to show that guilt, moreso than shame, motivate agents in social repair, restoration of relationships, and apologizing.[37] While guilt targets a specific wrong action, or as Hieronymi argues, the claim that the wronging leverages against the victim, shame can take as its object any fact about oneself. Both Ferguson et al. (1991), and Vaish et al. (2011), have shown that even children as young as 5-years-old are capable of understanding the appeasement functions of their own feelings of guilt. In "The Ethics of Guilt," William Neblett states that, "Ordinarily, we morally 'ought' to feel feelings of guilt over wrongdoings."[38] Not only is he correct in this assertion but as good moral agents we generally *do* feel guilt over actual moral transgressions. It is this type of unambiguous and sincere guilt that I propose is the type of attitude or mental state required of wrongdoers in successful apologizing. Much like the forswearing of resentment and the affect that accompanies such a change, the sincere affect required of wrongdoers here might also be at least partially involuntary. Indeed, much like a sincere change in affect in forgiving, a sincere attitude of guilt might also take time.

Further, this condition holds that it must also be the case that the wrongdoer is apologizing for *the right kinds of reasons*. What, exactly, this means becomes clearer if we look, again, to Neblett. Neblett states that, "Morality makes it incumbent upon us to feel guilt, and morality

[36] It may appear counterintuitive that the allocution requirement will address this question of sincerity. You might be thinking, why not the sincerity condition? While the allocution condition gives us insight into the wrongdoer's understanding and confession of sincere guilt, the sincerity condition makes it possible for wrongdoer's to sincerely believe she can fulfill her promise to do better, while knowing that there are evidential reasons that she might fail.
[37] Allport, 1979; Amodio et al., 2007; Baumeister, et al., 1994; Baumeister et al., 1995; Ferguson et al., 1991; Hacker, 2017, Jankowski and Takahashi, 2014; Kaufman 2011; Muris et al., 2013; Neblett 1974; Nussbaum 2005; Robertson et al., 2018; Tangney and Dearing, 2002; Vaish et al., 2011.
[38] Neblett (1974) pp.653

provides warranted ways for our feelings of guilt to be "discharged," i.e., it provides and permits us to redeem ourselves."[39] This is precisely where reasons can enter the conversation about guilt and its restorative aims.

In successful apologizing, it is clear that the wrongdoer cannot merely apologize for the purpose of discharging his or her uncomfortable feelings of guilt. An acceptable apology requires that we go much further than this. While a wrongdoer here might have the correct mental state of unambiguous guilt, and he may appear to be apologizing by doing the right sorts of things, he must apologize for the right kinds of reasons as well—i.e., because he wants to restore the relationship, or make things right in whatever way possible with the victim of his wronging. Again, as good moral agents, we cannot simply apologize for the purpose of discharging our feelings of guilt, we must also apologize with an attitude of care—one that ensures that *we have no more reason to feel guilty*. I posit this attitudinal/affective condition in the hopes that it will make it such that good apologies cannot obtain or be rendered successful in cases where the wrongdoer is apologizing for the wrong kinds of reasons. In the types of cases of this type, where the wrongdoer is apologizing for the wrong kinds of reasons, and the victim refuses to *accept* the apology, either one of two things is occurring:

Either:

> a. the correctness conditions for a successful apology have not been met on the wrongdoer's part,
>
> or,
>
> b. if the success conditions for a good apology have been met, then the victim is not doing what might be the morally right thing, which would be to accept the apology and move on.

[39] Ibid. pp.656

It strikes me that in this type of case, where a wrongdoer is apologizing merely to discharge their uncomfortable feelings of guilt, that they are apologizing for the wrong kinds of reasons. So, we should conclude here that (a) the wrongdoer did not actually meet, at least, this particular success condition for an acceptable apology, namely, the attitudinal / affective condition.

4.3: The Sincerity Condition:

In positing what makes a good apology, we must also consider what makes an apology sincere. The wrongdoer must express a sincere desire for restoration of the individual relationship, or at the very least to repair the harm she has caused if possible, and the wrongdoer must *sincerely believe* that she can follow through with this claim and whatever it entails. If the wrongdoer does not sincerely believe that she can ϕ, or that there is a significant chance that she cannot follow through on the change in behavior that she promises in apologizing, then there is a question as to whether she can still be said to sincerely apologize.

The sincerity condition I posit here is one of the most important conditions for good apologies. However, it is also one that needs perhaps the most explanation. The sincerity condition proposes that insofar as an apology can be said to be like a promise of better behavior, one should be able to reasonably believe that one can hold up his or her end of the promise made in apologizing. To motivate this condition, I will appeal to Berislav Marušić's paper, "Promising Against the Evidence." Marušić notes that in promising we face an issue that arises out of evidentialist accounts of belief and reasons for belief. If one cannot reasonably and sincerely believe that one will follow through on the promises she makes, then how can one responsibly make such promises? It seems like in our promissory apologies of better behavior, we face a similar problem.

If I believe that there is a good chance that I will fail in upholding my end of a promissory apology as a wrongdoer, can I sincerely apologize? Further, how can the victim *trust* that the wrongdoer will uphold her end of the deal when the victim has good evidence (the wrongdoing itself!) that the wrongdoer is not only capable of such things, but has enacted them in the past? For the purposes of space, I will not be delving too deeply into the literature about evidentialism and promissory obligation here. I will just briefly state that in many cases of promising against the evidence, if we only take an evidentialist account of intentions, as Marušić claims, we end up either making promises that will leave the promisee disappointed or indignant. Alternatively, such promises made for purely evidentialist reasons would render the promisor irrational or irresponsible. Ultimately, Marušić comes down on the evidentialist view and argues that one should not look to the evidence *alone* to predict what one will or will not do in promise keeping. Doing so treats the question as though it is not up to the agent—but it is! Marušić explains that to appeal to the evidence that one is unlikely to be able to keep her promise is to undermine one's agency. In cases where there is an uncertain future, as agents we, "should not seek to predict what we will do; we should engage in practical reasoning, not theoretical reasoning."[40] Marušić gives an example of an individual whose spouse goes off to war for an indefinite period of time. The individual in the example faces the decision of whether or not to promise that she will stay faithful to her spouse as her spouse leaves for the war. The individual wants to make such a promise even though she knows that there is a good chance she may never see her spouse again, her spouse may die in combat, or that her spouse may be gone for so long that the individual might waver in her ability to stay faithful and thus fails. Ultimately, Marušić argues that it would be irrational not to consider our evidence in cases of practical reasoning, but the evidence alone is sufficient to settle

[40] Marušić 2013, pp.308-309.

the question of what to do. Considering the evidence can tell us whether one course of action will be more or less difficult, and therefore more or less attractive to us. However, Marušić also argues that, "If you look to your evidential reasons alone, you do not give proper weight to your practical reasons—to what you value, want, or morally ought to do. In that way, you deny or distort your agency."[41] Essentially, when it is up to us as agents to do something, we can rationally believe we will do it, even when there is significant evidence to the contrary, "provided that it is practically rational for us to do it […]."[42] Marušić contends that in the case where someone is in a position to decide whether or not to wait for her spouse to return from the war,[43] the question can and should be settled by *deciding* to wait. Not only this, but such a decision should be made on both practical and evidential reasons, or again to use Hieronymi's terminology, such a decision should be made on the *right kinds of reasons*.[44] So, much like the promisor, the wrongdoer may be able to sincerely and rationally apologize or make a promise of better behavior against the evidence. This makes it such that the wrongdoer can sincerely apologize and can sincerely *believe* that they will be able to follow through on their promise of better behavior despite having evidence (e.g., their past actions) that they will fail.

This still leaves open the question as to whether the victim can *rationally trust* whether the wrongdoer will keep her promise in apologizing or not. If the forswearing of resentment is required in the process of completing an apology, this seems like a much more daunting task than if it is not. As I will show in the following, the distinction between acceptance and forgiveness will make it such that the victim can withhold her judgement about whether to forgive. At the time of the apology, assuming it is one that is sincere and meets the other proposed conditions for

[41] Ibid. pp.310
[42] Ibid. pp.311
[43] And in other cases like this—those of promising against the evidence.
[44] Hieronymi 2005

acceptability, the victim can choose to trust that the wrongdoer will *decide* (or at least sincerely attempt) to do the things they set out to do. Acceptance of an apology then, might be akin to choosing to trust, and then given further evidence at a later time, the victim can decide whether or not to forswear their resentment and forgive.

All of this being said, the sincerity condition holds that a successful apology requires that the wrongdoer *sincerely believe* that she can hold up her end of the deal in her promissory apology. The fact that every apology that entails a promise of better behavior, also entails a promise against the evidence is not a problem for this condition as it should now be clear that one can sincerely make such promises, and therefore such apologies.

4.4: The Allocution Condition:

In successfully apologizing, the wrongdoer must express her understanding and her reasoning for her understanding of what exactly she has done wrong. There is an additional element of generating a foundation of common belief posited by this condition.

This condition may seem odd at first glance but in many cases, it can make what might have been a bad apology acceptable. Allocution, in American legal theory is essentially the opportunity given to a defendant who has already pled guilty, to make a statement for herself. Often in legal cases the process of allocution is meant to have a humanizing effect for the defendant, and in some cases can result in a change in sentencing. In an allocution statement, the guilty party explains her reasoning process that led her to enact whatever crime she has committed. Additionally, it gives the defendant an opportunity to publicly apologize to the victim, the judge, the state, or what have you. I intend to appropriate this term for my purposes here. While I do not mean to imply that wrongdoers should go through the literal legal process of allocution in apologizing, it seems apt to apply this idea in this context. In legal contexts, allocution only occurs

once the defendant has already pled guilty, and similarly my view here covers apologies wherein the wrongdoer has committed some actual transgression, knows that they have done so, and feels unambiguous guilt. Legally and interpersonally then, allocution can have a validating effect for both the wrongdoer who feels sincere guilt, and the victim for whom the allocution process should affirm that the wrong leveraged against them was *actually wrong and/or harmful*. I contend that in successful apologizing, the wrongdoer may go through some such allocution process. In legal contexts, allocution is necessarily a written document or a speech act. In this way, allocution itself could satisfy my condition (1): the performativity condition. Allocution can be as minimal as the wrongdoer simply explaining her reasons for why she has transgressed as she has, or it may be a longer explanation that does something like affirming the harmfulness of what she has done to the victim. Engaging in an act of allocution as a part of an apology is often quite morally powerful, but this alone cannot render an apology acceptable: the other conditions need to be met as well.

A very good example of a case of allocution in successful apologizing is the case regarding Dan Harmon and Megan Ganz. Dan Harmon is a TV show runner and comedy writer who led such projects as the sitcom "Community," and the movie, "Monster House." Harmon faced sexual harassment allegations from one of his writers on the show, "Community," Megan Ganz. Harmon denied these claims for an entire season before accepting that he had in fact done something wrong and attempting to apologize to her. Ganz did not accept this first attempt at an apology: the apology was sent on Twitter, and was very brief in nature, essentially stating no more than "I regret that I did [x, y, and z]," or something like this. Ganz did not accept this apology, and in fact, specifically asked Harmon to go through the process of allocution: to explain to her what his reasoning was in his actions, and to effectively illustrate that he understood what he had done wrong, and why these acts were wrong. Years later, Harmon gave a very long public apology that ultimately Ganz did

accept. Harmon's successful apology, I think, can be held up as a paradigm example of allocution in apologizing. The following is a (long) excerpt from the transcription of Harmon's apology to Ganz:

> In 2000-whatever-whatever […] I had the privilege of running a network sitcom and I was attracted to a[n] employee. […] I mean, the most clinical way I can put it in fessing up to my crimes is that I was attracted to a writer that I had power over because I was a showrunner, and I knew enough to know these feelings were bad news. That was easy enough to know. I knew that they ran the risk of undercutting people's faith in my judgment, her faith in her talent, the other writers' respect for me, the entire production, the audience. […] And so I did the cowardly easiest laziest thing you could do with feelings like that and I didn't deal with them, and in not dealing with them, I made everybody else deal with them – especially her. Flirty, creepy, everything other than overt enough to constitute betraying your live-in girlfriend to whom you're going home every night, who is actually smart enough and respectful enough to ask you, "Do you have feelings for that young writer that you're talking about, that you're paying all this attention to?" And saying to her, "No," because the trick is, if you lie to yourself you can lie to everybody, it's really easy. […] And so I let myself keep doing it. And it's not as if this person didn't repeatedly communicate to me the idea that what I was doing was divesting her of a recourse to integrity. […] And so, after a season of playing it that way, I broke up with my girlfriend […] because I thought that would make having inappropriate feelings for a coworker appropriate if I wasn't involved. […] I broke up with my girlfriend, then I went right full-steam into creeping on my employee now it was even less inappropriate after all. Now I wasn't in danger of being a bad person. […] I crushed on her and resented her for not reciprocating it. And the entire time, I was the one writing her paychecks and in control of whether she stayed or went and whether she felt good about herself or not, and said horrible things, just treated her cruelly, pointedly. Things that I would never ever ever have done if she had been male and if I had never had those feelings for her. […] And I never did it before, and I will never do it again, but I certainly wouldn't have been able to do it if I had any respect for women on a fundamental level. I was thinking about them as different creatures, I was thinking about the ones that I liked as having some special role in my life, and I did it all by not thinking about it. So I just wanna say, in addition to obviously being sorry – but that's really not the important thing – I wanna say I did it by not thinking about it, and I got away with it by not thinking about it. And if she hadn't mentioned something on Twitter, I would've continued to not have to think about it, although I did walk around with my stomach in knots about it. But I wouldn't have had to talk about it. […][45]

While this is quite a long excerpt from the transcription of Harmon's apology, it shows the value of my proposed allocution condition. In reading Harmon's apology, one can see that he has fully considered the reasons for his actions, and what harms he has enacted. Not only this, but his

[45] *This American Life*, Episode "Get a Spine!" See the transcription of this episode for the full text of Harmon's apology.

statement also ensures that in receiving such an apology, Ganz understands that Harmon knows his actions were wrong, and that he knows why such things were wrongdoings. In this way, the allocution condition also ensures that the wrongdoer and the victim share a common belief about the transgression in question. This foundation of common belief is operative in making what might otherwise be a non-apology not only successful, but acceptable to the victim, in the sense that it can start the process of moral repair.

4.5: The Agent-Relative Condition:

Finally, if the victim has some personal requirement (within reason) for what would make an apology acceptable or valid for her, then the wrongdoer must also fulfill this. This final condition makes it possible for the victim to have a say in what features would make an apology acceptable specifically to her. By positing such a condition, I run the risk of proposing a potentially undermining factor. Meaning, this situational condition could undermine other good-making features of particular apologies. But this needn't be hugely problematic for my view. This condition can only function within reason, and there are certainly many types of demands that a victim could make that could fit the situational condition but render an apology morally invalid. To motivate this condition, I will appeal briefly to Jonathan Dancy's particularism about reasons, though the discussion of this condition needn't be purely metaethical in nature. To put it very briefly, Dancy's particularism about reasons—and specifically moral reasons—posits that something can be a reason for action for one person in one context, but that same thing may not be a reason for acting across all contexts for all people. This is to mean that a given feature of a scenario can be a reason for me to ϕ, but that same feature would not necessarily be a reason for you to ϕ. And in fact, for you this feature might be a reason to not ϕ. Dancy endorses the idea that moral reasons can have a sort of variable relevance given the different social facts of a situation.

The particularist view makes it such that the larger overarching moral landscape in which one finds oneself cannot be left out when we talk about reasons, and especially about moral reasoning. Additionally, Dancy gives us the idea of the valence switching of reasons: in one context something can be good, and in another context the very same thing can be bad. When applied to successful apologizing, this might be restated as: some feature that would make an apology acceptable for me, might make an apology unacceptable for you. Basically, this allows for the victim to make particular demands; they are empowered to ask for particular things that would make a perhaps otherwise not-so-good apology, good, for them.

Of course, these particular reasons must also be *reasonable*. I cannot reasonably demand that in apologizing to me, you must now also renounce your career and become an ascetic to truly understand the harm you've caused me. The types of particular features or claims that could make an apology acceptable to one person, but unacceptable to another, are features that should already be quite familiar to us. The situational or particular features that could make an apology acceptable for a given victim must be not only possible but also reasonable for the wrongdoer to incorporate in her apology. I have in mind here, features like (4) the allocution condition. I can imagine a case of some wrongdoer beginning the process of allocution, explaining her reasons and motivations and her understanding of the harm she has caused. The victim may quite fairly object, "I don't want to hear your reasons! Don't waste my time with this. Simply show me that you can do better!" In this type of case, the victim can claim that the omission of, say, allocution, might make the wrongdoer's sincere apology acceptable for her, when it otherwise would not be. On the other hand, if we think back to the example given in the previous section 3.4 on allocution, it was Ganz' explicit demand that Harmon go through the process of allocution— in Harmon's first attempt at apologizing to Ganz he did *not* do so, and she contacted him asking him to think about the

wrongdoing more, and give more explanation and reasons and essentially try again— and this feature is precisely what *would* make an apology from Harmon acceptable to Ganz. This illustrates the particularist point, and the motivation for positing this agent-relative condition quite nicely. If there is some such reasonable feature that would render an apology acceptable to a victim, then these features should also be met by the wrongdoer in apologizing.

Now, there is a lot that rests on the idea of the reasonability of these types of particular claims. Not only this, but this condition also presumes that the victim can somehow comfortably and safely communicate to the wrongdoer what her particular acceptability features of an apology are. There could very well be cases where the victim desires some specific feature of an apology that would make it acceptable, but wherein she does not feel safe communicating such a feature to the wrongdoer. It seems that in such cases, if the victim is reasonably fearful that she might be in some danger or be victim to further wrongdoing by communicating such a feature, then it is also likely that the wrongdoer is not in a position to give an acceptable apology as measured by all of the other conditions (1-4). In these cases, even if the wrongdoer is in a position to give an acceptable apology; even if the wrongdoer feels sincere guilt, and has reasoned about her wrong actions, and has the sincere desire to do better, if the victim is rationally fearful that she might be threatened, coerced, or further harmed in some way, then this alone would render the apology morally invalid.

My positive proposal for the situational or particular condition is simply to make it possible for a good apology to have agent-relative value. That is, if the victim can enact some reasonable feature that will make her wrongdoer's apology acceptable to her, when it would not be otherwise, then she should be able to make such a claim.

Section 5: Conclusion:

I have argued here for an account of what successful apologizing entails. The wrongdoer must (1) make her apology as a speech act, she must (2) have the right attitude in doing so and she must do so for the right reasons. The wrongdoer must also (3) sincerely believe that she will be able to follow through on the things that she promises in apologizing. The wrongdoer may (4) engage in the process of allocution, and in doing so help create a foundation of common belief. And finally, the wrongdoer may (5) incorporate some particular feature(s) that the victim claims would make an otherwise unacceptable apology acceptable.

I have also argued for an imperative distinction between acceptance of an apology and the forgiveness of a wrongdoer. Acceptance, in lieu of forgiveness then, makes it possible for the victim to essentially "wait and see" whether the wrongdoer will hold up her end of her promissory apology. Acceptance makes it possible that the victim is never required to forswear resentment in accepting an apology that promises better behavior—she can instead choose to trust that the wrongdoer will follow through on what is promised in said apology and reappraise her warranted resentment later in light of further actions and evidence. In fact, one could even conceive of cases where a wrongdoer has transgressed in some unforgiveable way, but perhaps she still offers an acceptable apology to her victim. The victim, in such a case, can rationally accept such an apology, but she needn't. However, it is important to note here that there is a choice the victim can make in such cases: the choice of whether to trust the wrongdoer. Indeed, there could be some acts that are simply unforgiveable. The victim might reasonably hold onto her resentment forever in such cases, but perhaps she accepts a sincere apology simply to allow herself and the wrongdoer to move on.

Acceptance then, bears a lot more moral weight than one might initially think. Acceptance can, in some cases, be akin to willingness to trust. In such cases the victim accepts the wrongdoer's sincere apology and renders the process of apologizing complete. However, after the completion of the apology, the victim can still rationally feel resentment, and perhaps wait and see if the wrongdoer really does hold up her end of the promissory apology. The victim needn't immediately forgive the wrongdoer just because a successful apology has taken place: the addition of acceptance to the moral schema gives us grades of supererogation. In accepting and completing a sincere apology, the victim cannot be found blameworthy or irrational for still experiencing feelings of resentment after this time. Perhaps the victim still feels resentment until she sees that the wrongdoer really has changed her behavior in the ways promised. Perhaps at this point, after the restoration process has taken place in some way or another, the victim's resentment will be alleviated, and she will have reason to forgive.

Bibliography

Abramson, Kate (2014). "Turning up the lights on gaslighting." *Philosophical Perspectives* 28 (1):1-30.

————, (2024). *On Gaslighting.* Princeton: Princeton University Press.

Alexander, Larry. "The Moral Magic of Consent (II)." *Legal Theory* 2, no. 3 (1996): 165–74. doi:10.1017/S1352325200000471.

Allais, L. "Feeling Blame and Feeling Forgiveness." Pettigrove, Glen & Enright, Robert (eds.) (2023). The Routledge Handbook of the Philosophy and Psychology of Forgiveness. Routledge.

Allport, G. W. (1979). *The Nature of Prejudice* (Unabridged, 25th anniversary). Addison-Wesley Pub.

Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2007). A Dynamic Model of Guilt: Implications for Motivation and Self-Regulation in the Context of Prejudice. *Psychological Science*, 18(6), 524–530. http://www.jstor.org/stable/40064652

Andersen, S. M., & Chen, S. (2002). "The relational self: An interpersonal social-cognitive theory." *Psychological Review*, 109(4), 619–645. https://doi.org/10.1037/0033-295X.109.4.619

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An interpersonal approach. *Psychological Bulletin*, 115(2), 243–267. https://doi.org/10.1037/0033-2909.115.2.243

————, (1995) Personal Narratives About Guilt: Role in Action Control and Interpersonal Relationships. *Basic and Applied Social Psychology*, 17(1-2), 173-198, DOI: 10.1080/01973533.1995.9646138

Calhoun, Cheshire. "Changing One's Heart." Ethics 103, no. 1 (1992): 76–96.

http://www.jstor.org/stable/2381496.

Call, J., & Tomasello, M. (1999). A nonverbal false belief task: The performance of children and great apes. Child Development, 70(2), 381–395. https://doi.org/10.1111/1467-8624.00028

Chudek M, Henrich J. Culture-gene coevolution, norm-psychology and the emergence of human prosociality. Trends Cogn Sci. 2011 May;15(5):218-26. doi: 10.1016/j.tics.2011.03.003. Epub 2011 Apr 7. PMID: 21482176.

Dancy, Jonathan. "The Particularist's Progress." *In Recent Work on Intrinsic Value*, pp. 325-347. Springer, Dordrecht, 2005.

Deweese-Boyd, Ian (2023). Self-deception. *Stanford Encyclopedia of Philosophy*.

Direct-Caterpillar77. March 9, 2018. "I [27 F] think my boyfriend [29 M] booby-trapped our apartment, I found something and don't know how to bring it up." [Online forum post] Reddit. https://www.reddit.com/r/BestofRedditorUpdates/comments/1ivb2gl/i_27_f_think_my_boyfriend_29_m_boobytrapped_our/

Duff, R. A., 1986, Trials and Punishments. Cambridge: Cambridge University Press.

Ferguson, T. J., Stegge, H., & Damhuis, I. (1991). "Children's understanding of guilt and shame." *Child Development*, 62(4), 827-839.

Fricker, Miranda (2007). *Epistemic injustice: power and the ethics of knowing*. New York: Oxford University Press.

Funkhouser, Eric (2019). *Self-deception*. New York, NY: Routledge.

Glass, Ira. (Host). (2019, May 10). "Get a Spine!" no.673, act 1. In *This American Life*. NPR. https://www.thisamericanlife.org/674/get-a-spine

Gunn, Hanna ; Longair, Holly & Oliver, Kelly (eds.) (2025). *Gaslighting: Philosophical*

Approaches. New York: State University of New York Press.

Hacker, P. (2017). Shame, embarrassment, and guilt. *Midwest Studies in Philosophy*, 41, 202-224.

Harman, Elizabeth. "Morally Permissible Moral Mistakes" *Ethics* 126, no.2 (2016): 366-393. https://doi.org/10.1086/683539

Hieronymi, Pamela. "Articulating an Uncompromising Forgiveness." *Philosophy and Phenomenological Research* 62, no. 3 (2001): 529–55. https://doi.org/10.2307/2653535.

————, "The Wrong Kind of Reason." *The Journal of Philosophy* 102, no. 9 (2005): 437–57. http://www.jstor.org/stable/3655632.

————, 2004, "The Force and Fairness of Blame", Philosophical Perspectives, 18: 115–148.

Jankowski KF, Takahashi H. "Cognitive neuroscience of social emotions and implications for psychopathology: examining embarrassment, guilt, envy, and schadenfreude." *Psychiatry Clin Neurosci*. 2014 May;68(5):319-36. doi: 10.1111/pcn.12182. PMID: 24649887.

Kaufman, W. (2011). Understanding honor: beyond the shame/guilt dichotomy. *Social Theory and Practice*, 37(4), 557-573.

Kelly, D., & Westra, E. (2026, February 27), "The Psychology of Articulated Norms," Contributed Symposium: The Cognitive Science of Ambivalent Norms. Southern Society of Philosophy and Psychology, Atlanta, GA.

Kelly, Daniel, Evan Westra, and Stephen Setman, "The Psychology of Normative Cognition", *The Stanford Encyclopedia of Philosophy* (Spring 2025 Edition), Edward N. Zalta & Uri Nodelman (eds.), forthcoming URL = <https://plato.stanford.edu/archives/spr2025/entries/psychology-normative-cognition/>.

Klein, W., Wood, S., & Bartz, J. A. (2025). A Theoretical Framework for Studying the

Phenomenon of Gaslighting. *Personality and Social Psychology Review*, 0(0). https://doi.org/10.1177/10888683251342291

Langton, Rae (1993). "Speech acts and unspeakable acts." *Philosophy and Public Affairs* 22 (4):293-330.

Liberto, Hallie. "The Problem with Sexual Promises." *Ethics* 127, no. 2 (2017): 383-414.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. Psychological Inquiry, 25(2), 147–186. https://doi.org/10.1080/1047840X.2014.877340

Martin, Adrienne M. "Owning up and Lowering Down: the power of apology." *The Journal of Philosophy* 107, no. 10 (2010): 534–53. http://www.jstor.org/stable/29778053.

Marušić, Berislav. "Promising against the Evidence." *Ethics* 123, no. 2 (2013): 292–317. https://doi.org/10.1086/668704.

McGeer, V., 2013, "Civilizing Blame", in Coates & Tognazzini 2013: 162–188.

McGregor, Joan. "Consent and Autonomy," *Is it Rape?* Routledge. (2005).

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

————, 2013, "Directed Blame and Conversation", in Coates & Tognazzini 2013: 119–140.

Mele, Alfred R. (2001) *Self-Deception Unmasked.* Princeton University Press.

————, (1997). Real Self-Deception. *Behavioral and Brain Sciences* 20 (1):91-102.

————, (1983). Self-deception. *Philosophical Quarterly* 33 (October):366-377.

Metro-Goldwyn-Mayer Corp. (1944). *Gaslight* [Film]. Directed by George Cukor.

Moran, K. 'Moving on for Community's Sake: A Self-Respecting Kantian Account of Forgiveness' in Stefano Bacin, Alfredo Ferrarin, Claudio La Rocca Margit Ruffing (eds.) *Kant und die Philosophie in Weltbürgerlicher Absicht: Proceedings of the XI International Kant Congress*, De Gruyter, 2013.

Muris, P., Meesters, C., Cima, M., Verhagen, M., Brochard, N., Sanders, A., Meesters, V. (2013). "Bound to Feel Bad About Oneself: Relations Between attachment and the Self-conscious Emotions of Guilt and Shame in Children and Adolescents." *Journal of Child and Family Studies*, 23(7), 1278–1288. https://doi.org/10.1007/S10826-013-9817-Z

Neblett, William (1974). "The Ethics of Guilt." *Journal of Philosophy* no.71 (18):652-663.

Nelkin, Dana K. (2002). Self-deception, motivation, and the desire to believe. *Pacific Philosophical Quarterly* 83 (4):384-406.

Nussbaum, Martha. "Inscribing the Face: Shame, Stigma and Punishment." *Nomos* 46 (2005): 259–302. http://www.jstor.org/stable/24220152.

Pickard H. Responsibility Without Blame: Empathy and the Effective Treatment of Personality Disorder. Philos Psychiatr Psychol. 2011 Sep;18(3):209-223. doi: 10.1353/ppp.2011.0032. PMID: 22318087; PMCID: PMC3272423.

Robertson, T. E., Sznycer, D., Delton, A. W., Tooby, J., & Cosmides, L. (2018). The true trigger of shame: Social devaluation is sufficient, wrongdoing is unnecessary. *Evolution and Human Behavior*, 39(5), 566-573.

Scanlon, TM. *What We Owe to Each Other.* Harvard University Press. 1998.

Shoemaker, David & Vargas, Manuel (2019). Moral torch fishing: A signaling theory of blame. Noûs (3).

Smith, A. M. (2012). Attributability, Answerability, and Accountability: In Defense of a Unified Account. Ethics, 122(3), 575–589. https://doi.org/10.1086/664752

Stoljar, Natalie, "Feminist Perspectives on Autonomy", The Stanford Encyclopedia of Philosophy (Summer 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/sum2024/entries/feminism-autonomy/>.

Talbert, M., 2012, “Moral Competence, Moral Blame, and Protest”, The Journal of Ethics, 16: 89–109.

Tangney, J. P., & Dearing, R. L. *Shame and Guilt.* New York: Guilford Press. (2002).

Tognazzini, Neal and D. Justin Coates, "Blame", The Stanford Encyclopedia of Philosophy (Summer 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2021/entries/blame/>.

Trivers, R. (2011). *The folly of fools: The logic of deceit and self-deception in human life*. Basic Books.

————, (2000/2002). “Self-Deception in Service of Deceit.” Reprinted in *Natural Selection and Social Theory*. Oxford University Press.

Vaish, A., Carpenter, M., & Tomasello, M. (2011). “Young children's responses to guilt displays.” *Developmental Psychology*, 47(5), 1248.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition, 13(1), 103–128. https://doi.org/10.1016/0010-0277(83)90004-5