UNIVERSITY OF CALIFORNIA SAN DIEGO

A Theoretical and Empirical Investigation of Prosocial Lying

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Management

by

Matthew James Lupoli

Committee in Charge:

Professor Christopher Oveis, Chair
Professor James Andreoni
Professor Ayelet Gneezy
Professor Yuval Rottenstreich
Professor Pamela Smith

The Dissertation of Matthew James Lupoli is approved,
and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2018

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Chapter 2, in full, is a reprint of material as it appears in Journal of Experimental Psychology: General, which was co-authored by Lily Jampol and Christopher Oveis in 2017. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of material as it appears in Organizational Behavior and Human Decision Processes, which was co-authored by Emma E. Levine and Adam Eric Greenberg in 2018. The dissertation author was the primary investigator and author of this paper.

VITA

EDUCATION

University of California San Diego Rady School of Management
Ph.D. in Management, Expected June 2018

Sarah Lawrence College
B.A. in Psychology, May 2009

RESEARCH INTERESTS

Prosocial Behavior, Deception, Field Experiments, Ethics, Emotion, Choice Architecture

PUBLICATIONS

Lupoli, M.J. (2018) Lying because we care. *Forthcoming at Rady Business Journal*.

Lupoli, M.J., Levine, E.E., Greenberg, A.E. Paternalistic lies (2018). *Organizational Behavior and Human Decision Processes*, *146*, 31-50.

Lupoli, M.J., Oveis, C., Jampol, L.E. (2017). Lying because we care: Compassion increases prosocial lying. *Journal of Experimental Psychology: General*, *146*(7), 1026-1042.

ABSTRACT OF THE DISSERTATION

A Theoretical and Empirical Investigation of Prosocial Lying

by

Matthew James Lupoli

Doctor of Philosophy in Management

University of California San Diego, 2018

Professor Christopher Oveis, Chair

Abstract

Prosocial lying, or lying that is intended to benefit others, is a ubiquitous phenomenon that can have profound consequences. Despite their prevalence and importance in social life, little research has investigated the causes and downstream effects of these lies. In this dissertation, I first define prosocial lies and explain how they fit into the theoretical framework of lying and deception. I also present empirical evidence for a causal driver of prosocial lies, and identify a key determinant of whether these lies are viewed favorably or unfavorably by their recipients.

Despite honesty being widely held a virtue, dishonesty is prevalent in everyday life. The fact that honesty is considered to be a normative and valued behavior is understandable given the potentially harmful effects dishonesty can have on interpersonal relationships, groups, and nations. Yet, individuals frequently tell lies that are motivated not to help themselves at the expense of others, but rather by a desire to benefit others. To date, little research has examined the factors that drive individuals to tell these lies, or how people respond to them. In this dissertation, I examine antecedents and consequences of lies that are intended to help others.

This dissertation is organized as follows: In Chapter 1, I define prosocial lies and explain where these lies fit into the theoretical framework of lying and deception. I give special attention to the idea that intentions and consequences are a key dimensions by which prosocial lies can be defined, and illustrate the implications of this notion for our understanding of prosocial lying. I also provide a brief overview of the methods by which prosocial lies have been studied, and offer insights on how they might be examined in the future.

In Chapter 2, I present evidence for a causal driver of prosocial lying: compassion. The contents of this chapter have been published in *Journal of Experimental Psychology: General* (Lupoli, Jampol, & Oveis, 2017). In addition to the published contents, I also include a preface to this chapter that discusses emotion as a starting point for this project, as well as the role of this paper in a larger debate on the merits and disadvantages of empathy and compassion as harbingers of social good.

In Chapter 3, I examine how individuals respond when they learn that they have been the target of a prosocial lie. Specifically, I introduce the constructs of paternalistic lies and unequivocal prosocial lies—two subsets of prosocial lies—and explore how responses to these two classes of lies differ and why. The contents of this chapter were recently published in

Organizational Behavior and Human Decision Processes (Lupoli, Levine, & Greenberg, 2018). Similarly to Chapter 2, in this chapter I also include a preface to the published work. This introduction bridges conceptual work in Chapter 1 with empirical research presented in Chapter 3 by highlighting the latter's hybrid intention- and consequence-based perspective on prosocial lying.

In addition to summarizing the organization of this dissertation, it may be helpful to also provide some discussion of what this dissertation will contain, as well as what it will not contain.

First, this dissertation is not a comprehensive review of deception research. The aim here is not to describe in detail everything we know about deception, and there are some areas of research on deception that I will not be covering. Specifically, I do not discuss work on deception detection. This is a fascinating topic that has amassed a considerable body of work (e.g. Buller & Burgoon, 1996; DePaulo et al., 2003; Ekman & Friesen, 1969, 1974; Ekman & O'Sullivan, 1991; Porter & ten Brinke, 2008; ten Brinke, Stimson, & Carney, 2014). However, I view the questions that deception detection research strives to answer—that is, how we come to suspect and know whether a person is lying, and how the deceiver might successfully or unsuccessfully conceal this information—as fundamentally distinct from the questions I find most interesting, which are: What drives people to tell prosocial lies, and what are the consequences of these lies?

That said, I will be discussing approaches to the study of deception and prosocial lying. There are several dimensions on which different types of deception can be defined or classified. For example, classification can be based on the medium of deception (e.g., lying, paltering, deceptive omission, etc.), the content of the deception (e.g., what is being lied about), the intentions behind deception, and the consequences of deception. Addressing the dimensions on

which deception is defined is important for the current purposes because prosocial lying is necessarily classified based on one or two dimensions (i.e., intentions and/or consequences) but can take shape in many other "types" of deception as defined by other classifiers (e.g., prosocial lies of omission, as classified by the communicative form of deception). I will discuss this further in Chapter 1.

Another subject this dissertation does not address is whether lying is inherently immoral. While the morality of deception has long been debated by philosophers (Bok, 1978; Kant, 1785), I approach the study of prosocial lying with the foundational assumption (which is empirically verified, as described in Chapters 2 and 3) that regardless of deontological views about the morality of deception, people do tell (prosocial) lies.

Although I do not answer the question of whether "lying is wrong" is a correct moral rule, this dissertation does focus on moral judgments of prosocial lies in Chapter 3. As I discuss in that chapter as well as in Chapter 2, a key component of prosocial lying (and an important reason why I find these lies interesting) is that they represent a conflict between two moral values: honesty and kindness. An underlying theme of this dissertation concerns understanding how individuals navigate that tension. This includes determining when and why people tell prosocial lies, as well as how people view prosocial lies and those who tell them. Thus, while I do not say whether prosocial lying is immoral, I do offer theory and evidence as to when and why individuals view prosocial lying as reprehensible, and also discuss moderators of these views. Additionally, I provide some prescriptive advice based on my research about when prosocial lies are likely to have positive or negative effects on others. However, my recommendations are not sufficiently comprehensive to describe whether one should tell a prosocial lie across all situations in which this opportunity should arise.

Chapter 1

The primary goal of this chapter is to answer the questions: What is prosocial lying, and how do we measure it? I define prosocial lies as false statements that are made to intentionally, knowingly, and/or purposely mislead and benefit others.[1] On a surface level, this definition seems straightforward. However, to wholly comprehend the meaning of prosocial lying, one must also have an understanding of the constructs it is related to, as well as those from which it is different. For instance, what is lying and how does it compare to deception? What are the dimensions by which we differentiate types of lies? And how are prosocial lies similar or different compared to other types of lies? One must also understand the meaning of the word "prosocial" to understand prosocial lying. Specifically, what does it mean for a lie (or any behavior) to be prosocial? Are prosocial lies defined by intentions or consequences or both? How does the presence of mixed motives and undesirable results of these lies influence the degree to which they may be considered prosocial? In this chapter, I will answer these questions by drawing on theoretical and empirical work on both deception and prosocial behavior. By synthesizing these literatures, I fill an existing gap in the theoretical framework of prosocial lying.

What is Deception?

One cannot discuss lying without first mentioning deception. Deception is a higher order term for a broad class of deceptive behaviors that includes lying, deceptive omission, paltering,

[1] While others have defined prosocial lies as those that are intended to mislead and benefit others (Levine & Schweitzer, 2014), this definition of prosocial lying is the first to incorporate Levine's (2014) language of "intentionally, knowingly, and/or purposely."

and a host of other phenomena (more on this later; see Table 1.1 for definitions with examples). Deception is defined as the transmission of information to intentionally, knowingly, and/or purposely mislead others (Buller & Burgoon, 1996; Levine, 2014). There are four components of this definition that require elaboration to paint a complete picture of deception: (a) transmission of information; (b) intentionally, knowingly, and/or purposely; (c) mislead; and (d) others. Though I list these terms in the order in which they appear in the definition, I will discuss them in reverse order for narrative flow.

"Others". When I refer to deception, I really mean interpersonal deception. When interpersonal deception occurs, there is always a deceiver (the person(s) who deceives) and a target (the person(s) being deceived), who are not one and the same. Both the deceiver and target can be individuals or groups. For example, executives of an organization may collectively decide to lie to the public about uses of client data; or an individual might lie to the Internal Revenue Service about sources of income. According to this definition, deception cannot occur within a single individual.

Although I focus only on interpersonal deception in this paper, it is worth noting that there is a rich literature on positive illusions, which can be considered a type of self-deception. Positive illusions are characterized by inordinately positive self evaluations, inaccurate perception of control or mastery, and over-optimism (Taylor & Brown, 1988, 1994; Chance, Norton, Gino, & Ariely, 2011). As such, these illusions bear considerable theoretical differences compared with interpersonal deception. To name a few, interpersonal deception does not necessarily involve these manifestations of optimism, though positive illusions may play a role in motivating interpersonal deception (e.g., intentionally exaggerating one's ability to others to maintain positive self regard). Additionally, while positive illusions have documented positive

effects on mental and physical health (Taylor et al., 2000), deception may have positive or negative effects on the self and others (Boles, Croson, & Murnighan, 2000; Levine & Schweitzer, 2014, 2015; McCornack & Levine, 1990; Schweitzer, Hershey, & Bradlow, 2006; Tyler, Feldman, & Reichert, 2006). Although positive illusions are a fascinating and important topic of research, in light of these differences, I do not consider them to be a subset of (interpersonal) deception as I define it, and thus do not discuss these phenomena further.

"Mislead". Deception involves misleading others. The deceiver can do this either by instilling a false belief or by changing another's pre-existing belief to a false state (Knapp & Comadena, 1979; Zuckerman, DePaulo, & Rosenthal, 1981; Zuckerman & Driver, 1985). For example, if I tell an acquaintance that my degree is in medicine when it is actually in management, this is deceptive regardless of whether the acquaintance had a preconceived notion of what my degree was in (or even whether I was studying for one). To mislead others implies not only that a false belief is instilled in the recipient, but also that the deceiver knows or believes it to be false. In saying that the deceiver "believes it to be false," I mean that the deceiver does not necessarily need to have an accurate understanding of a state of the world in order to lie about that state. For instance, let's say your mother asks you if you called your brother for his birthday last week. You don't remember calling him, but you tell her that you did so that she doesn't get upset. In reality, however, you did call your brother, but don't remember doing so because you were preoccupied at the time and he did not answer the phone. In this case, you actually provided factual information to your mother, but still deceived her because you believed that information to be false. To go a step further, there need not be an objectively "true" state of the world for one to lie about that state. That is, one can deceive others about one's own feelings or opinions, which cannot be verified as true.

"Intentionally, knowingly, and/or purposely". That the deceiver has knowledge or beliefs that s/he has created a false belief in another is suggestive of the third essential characteristic of deception: intent. In order for an act to be considered deceptive, the deceiver must intentionally, knowingly, and/or purposely mislead another person (Levine, 2014). Deception is goal-directed. As a rule, people deceive when they perceive that the truth is in some way unacceptable, inconvenient, or undesirable. Thus, deception is used instrumentally with the intention or motivation to somehow eschew or distort that truth.

To illustrate the importance of intention in deception, consider cases where individuals make false statements that are *not* deceptive. For example, providing incorrect information that one does not know is incorrect is not deceptive. If I believe and tell you that it is Thursday when it is actually Friday, I am not lying because the false information is given accidentally and without the intention to mislead. Likewise, intentionally transparent lies or exaggerations (e.g., sarcasm) are not deceptive because the communicator does not intend the target to believe what is said as true. For example, if a friend asks me how I am doing after a tough day and I reply, "I'm dead," I am not lying because I expect that my friend will not believe what I say is true. Intention to mislead is a critical component of deception and without it, false statements are not deceptive.

To some readers, the words "intent" and "motivation" may suggest a premeditative or conscious component of deception. To assume that deception is a conscious decision is not unreasonable, as people can and do consciously weigh the costs and benefits before deciding whether to lie (e.g., Erat & Gneezy, 2012). However, the word "purposely" is included in the definition of deception to convey the idea that deception doesn't necessarily require conscious forethought. It is possible that an individual will become aware of or make inferences about the

reasons why s/he deceived only after engaging in deception (and perhaps not at all). For example, imagine that a woman suddenly asks her husband whether he has ever cheated on her. The husband has in fact cheated, but quickly replies that he has not. The man may reflect after the fact that by lying he aimed to avoid a serious argument, prevent his wife from feeling awful, and preserve the relationship. However, these specific thoughts may not have occurred to him in complete form in the roughly one second it took him to respond—a response that was based on an intuition that telling the truth was a bad idea. Although an analysis of the interplay between unconscious processes and behavior is beyond the scope of this paper, the notion that goal-oriented (i.e., purposeful) deception can occur without conscious planning is consistent with research showing that (a) beliefs, judgments, and emotions can occur outside of conscious awareness (Kunst-Wilson & Zajonc, 1980; Murphy & Zajonc, 1993; Zajonc, 1980); and (b) individuals often lack insight into the processes that influence their decisions (Bem, 1972; Nisbett & Wilson, 1977a, 1977b).

"Transmission of information". The facet of deception that has received arguably the most attention by deception researchers is the transmission of information—that is, how do people mislead others?

As the generality of the phrase implies, there are many ways that people can deceive or be deceived. A starting point to organizing types of deception and to understanding how these classes of deception relate to and differ from each other is to first consider the dimensions by which acts of deception can be classified. Deceptive behaviors can be classified based on several dimensions, including communicative method (e.g., how is the individual deceiving?), content of the deception (e.g., what is s/he lying about?), the relationship between the deceiver and target (e.g., who is lying and to whom?), and the intentions and consequences of deception (e.g., is the

lie prosocial or selfish?). As the term "prosocial lying" indicates, in this chapter I focus on two dimensions of deception in particular: the communicative method of deception (e.g., lying) and the intentions and consequences of deception (e.g., prosocial).

What is Lying? On Communicative Forms of Deception

Perhaps the most well-known classifier of deception is the communicative method by which people deceive or are deceived. Two main approaches have dominated research on communicative forms of deception. The first and most common perspective, which has been adapted in mostly disparate bodies of research in psychology (e.g., DePaulo et al., 1996), behavioral economics (e.g., Erat & Gneezy, 2012), and communication (e.g., Camden, Motley, & Wilson, 1984), is a discrete approach to deception. According to this framework, deception is the superordinate classification of a broad number of phenomena, or types of deception. These categories of deception are defined by differences in the communicative behaviors individuals use to mislead. Because communicative method is a popular classifier of deceptive behaviors, and because there have been inconsistencies in how classes of deception have been defined in the literature, it would be helpful to provide a rough map of these classes with precisely defined nomenclature. This taxonomy is not meant to be exhaustive in either the level of detail on each category of deception (this is not the focus of this dissertation) or the number of categories (which may be an impossible task, as discussed later). However, by describing types of deception at the construct level, I hopefully will provide some organization to a convoluted group of terms that prosocial lying and other deception researchers can implement going forward to maintain some standard of clarity and consistency.

One of the most commonly discussed methods of deception (which is also a key focus of this dissertation) is lying. Put simply, lies are defined as false statements that are made with the

intent to mislead (Boles, Croson & Murnighan, 2000; Gino & Shea, 2012; Levine & Schweitzer, 2014). One could also replace the phrase "the transmission of information" with the phrase "making false statements" in the definition of deception to arrive at the definition of lying (i.e., making false statements to intentionally, knowingly, and/or purposely mislead others). While lies are often verbal, they need not be; it is possible to lie through digital mediums, such as email, text messages, or multiple choice responses on a survey.

It is important to emphasize that lying involves making *false statements*, whereas it is possible to engage in deception in other ways without false statements. For instance, paltering is another communicative method of deception that involves making true statements that are aimed to mislead others (Rogers, Zeckhauser, Gino, Norton, & Schweitzer, 2017). One well-known example of paltering is when Bill Clinton said in an interview with Jim Lehrer that "there is no sexual relationship" with Monica Lewinsky in response to the question, "You had no sexual relationship with this young woman?" Despite previously having a sexual relationship with Lewinsky, Clinton technically made a true statement by using the present tense "is," since the relationship had ended before the time of the interview. However, the statement was intended to deceive, as the question asked if he *had* a sexual relationship, and many interpreted his statement to mean that he had not (example cited in Rogers et al., 2017).

Another communicative method of deception is deceptive omission, or failing to provide information with the intent to mislead. As far as I know, this is the first use of this terminology, as other work has labelled this phenomenon as lies of omission (e.g., Ekman & Friesen, 1974). However, deceptive omission is a more accurate label in this framework than lies of omission because lying involves actively making statements, whereas deceptive omission specifically involves *not* making statements.

One interesting question concerning deceptive omission is when does leaving out information constitute deception? Surely failing to expound the full extent of one's knowledge on a subject when asked a question about that subject is not necessarily deceptive. To illustrate, imagine John asks his wife Katie, "What did you to today?" Katie might respond by telling John about the delicious lunch she had, and about the latest run-in with her most-hated colleague. However, she might leave out other details of her day, such as flossing her teeth and having a glass of water. In this case, Katie did not deceive John because she had no intent to mislead; she merely shared information she thought he would be interested to hear, and withheld information that she assumed he would not care about or want to know. However, if Katie replied in that same way when she had also been fired from her job that day, this would be deceptive. By leaving out this tidbit, she would be withholding information that is essential to addressing the spirit of John's question. That is, John may not expect that Katie got fired, but if she did, this is information he would (a) want to know and (b) expect that she would tell him. Thus, to determine whether omitting information constitutes deception, one must consider whether the prospective deceiver violates the perceiver's sense of the target's expectations about what information should be shared. It would be interesting to test whether laypeople share this intuition, and when they believe an omission is deceptive, as these notions have not been examined empirically.

There are many other types of deception that differ by communicative method. Related to deceptive omission, minimizing the severity of a situation or problem can be considered deceptive because it leads the target to believe that something is less serious or requires less attention than is truly warranted. Conversely, exaggerations are also deceptive in that the target is led to believe that a situation or problem is more serious than warranted. Cheating is another

form of deception that involves covertly violating rules of conduct, where the target is either the purveyor of the task or game that is being cheated on, or others who are involved in that task or game. For instance, athletes who use illegal performance enhancing drugs are cheating their league and opposing teams, and those who move six spaces in Monopoly when they really rolled a five are cheating other players in the game. Disguises, camouflage, and illusions (that are not advertised as such) can all be considered acts of deception (but not lying) as well: in each case, the deceiver attempts to create the false impression in an observer that s/he is someone else, that s/he is not present, or that an illusion stemmed from magic or supernatural abilities and not natural means, respectively. Similarly to illusions, one can also deceive by misdirecting a person's attention or distracting them so as to prevent them from obtaining truthful information. In addition, communicating in code can be considered deceptive when the communications create a false impression in others. For instance, coded language (verbally, nonverbally, or digitally) can be used to deceive by having an agreed upon word or phrase that has one meaning to observers and another to colluders (e.g., encrypted language). It can also be used to disguise the fact that communication is actually occurring (e.g. hand signals).

As one can see, the number of communicative types of deception is large. However, aside from the unwieldy number communicative types of deception, coming up with an exhaustive list is difficult for two reasons. First, some forms of deceptive communication cannot easily be classified. Vague or ambiguous statements are a primary culprit. For example, imagine that a friend asks, "How are you feeling?" and you reply, "OK I guess," when in reality you feel dreadful. Granted the conversation goes no further, this statement can be considered deceptive in that it creates a false impression of your true internal state. But is it lying, deceptive omission,

minimization, or some combination of the three? Arguments can be made on multiple sides here.[2]

The second reason why a complete taxonomy of communicative forms of deception has not yet been established is that as technology advances, new ways to deceive others are emerging. Some computer-mediated types of deception can fit cleanly under existing labels. For example, using a false identity is a form of disguise. However, novel forms of computer-mediated deception are still being developed and are likely to materialize in the future. For instance, developers could create an app that replaces a user's location with a false location to prevent third-parties from observing the user's actual location. Deception of this kind cannot neatly be placed into existing buckets of communicative forms of deception because it does not resemble interpersonal contexts in which deception has been studied (i.e., before the technology existed).

Considering the problems with categorizing communicative forms of deception, some researchers have abandoned the discrete approach and attempted to provide a more parsimonious theoretical account of what deception is and is not. One of the most influential of these theories is McCornack's Information Manipulation Theory (IMT; McCornack, 1992; McCornack et al.,

[2] Note that a response of "fine, thanks" (or similar) to the question, "how are you" (rather than "how are you feeling") is arguably not a lie in the United States. In the United States, "how are you" is a platitude used as a greeting rather than an actual question. Thus, a person who responds positively to this greeting when they are actually feeling negatively may not be lying because they do not intend to mislead the greeter; the greeter has likely no expectation of a true or thoughtful response.

2014; McCornack, Levine, Solowczuk, Torres, & Campbell, 1992; Yeung, Levine, & Nishiyama, 1999). According to IMT, referring to communications as honest or dishonest oversimplifies deception because communications vary in the degree of dishonesty. McCornack argues that all deceptive communications can be explained by the extent to which they vary on at least one of four continuous dimensions: quantity (the amount of information delivered), quality (the extent to which information is distorted or fabricated), relation (the degree to which information provided is relevant), and clarity (the degree to which information is ambiguous).

IMT has several advantages over taxonomic approaches to deception. One advantage of is it creates a parsimonious model of deception that can account for several of the aforementioned classes of deception. For example, deceptive omission involves withholding information (quantity); lies involve providing false information (quality); misdirecting attention involves giving irrelevant information (relation); and vague statements are inherently ambiguous (clarity). Other instances of deception can be described by IMT using more than one dimension, whereas the same act(s) would be difficult to classify using the discrete approach. For example, imagine a man asks his partner, "what did you do last night?" and she replies "had some fun with friends at a party." If in reality she attended a party with friends but also slept with another man there, this would be a violation of quantity (leaving out information) and clarity (vaguely describing what happened at the party). Another advantage is that IMT does not simply dichotomize deceptive acts as either honest or dishonest, but rather accounts for a continuum of dishonesty on each dimension. For example, two different instances of deceptive omission may both be dishonest, but can vary in the amount of information that is left out. Thus, IMT allows us to more precisely describe acts of deception without creating an exaggerated dichotomy between honesty and dishonesty or putting incomplete or inaccurate labels on these acts.

However, IMT is not without problems. A first issue is that some of the dimensions are arguably not orthogonal. Specifically, while some acts of deception can be distinguished as to whether they involve a violation of quantity or clarity, others likely cannot. That is, some statements are ambiguous *because* they leave out information. As discussed earlier, the inability to classify certain deceptive acts in to categories is an issue with the taxonomic approach as well. However, the latter approach does have some advantages over IMT. For one, some classes of deception described previously cannot fit into the IMT framework. In particular, IMT does not account for nonverbal forms of deception, such as disguises and encrypted information. Another problem is that two communications that are equally deceptive on the same IMT dimension may be perceived differently by targets and observers. For instance, it is possible that lying to someone's face may be seen as a more severe moral violation than verbalizing that same lie over email. As new digital forms of deception emerge, it will be important to better understand how deceptive medium influences responses to deception. A final issue is one of practicality: While numerous, taxonomic classes of deception may be easier to conceptualize than the four-dimensional model. Should we call an intentionally false statement a lie or a quality of information violation? Researchers' classifications of deception stemmed at least in part from their lay understanding of how acts of deception qualitatively differ from each other. That understanding shares a common language that contains words like "lying" and "omission." To the extent that these classes more intuitively describe the experience of deception according to existing schemas, this system may make the study of deception easier.

What is Prosocial Lying? On the Importance of Intentions, Consequences, and Mixed Motives

While understanding how deception is communicated is essential to the study of deception, this understanding provides only a starting point to answering other questions about deception, such as: Why do people lie? And how do individuals judge lying, and those who lie? To address these questions, one must focus on another dimension by which deception can be defined—that is, the intentions and consequences of deception.

When we hear the word "lies", most of us probably have automatic negative connotations; we might recollect a time where we felt betrayed after being lied to, or about feeling guilty after telling a lie. Indeed, individuals cite honesty as an important moral value (Graham et al., 2015), and some people believe that any type of lying is wrong (Bok, 1978; Kant, 1785; Erat & Gneezy, 2012). But the fact remains that people do lie, and they do it quite regularly (DePaulo et al., 1996). This does not mean, however, that those who lie are necessarily depraved individuals who are only looking out for their self-interest. To fully understand what leads people to deceive and how these deceptive acts affect others, one must first examine people's motivations behind lying.

People lie for many reasons. They might lie for their own financial gain, to improve their reputation, to prevent others from feeling badly, to avoid punishment or embarrassment, or to preserve a relationship. These are just a few examples, and it would not be useful nor perhaps even possible to list all the reasons. Instead, a better approach would be to understand what are the broad classifiers of motivations behind deception. By arriving at a more parsimonious classification system for these motivations, we can hopefully draw conclusions about a wide variety of contexts in which deception occurs using simplified experimental designs that model these contexts. One classifier that has substantial explanatory power for understanding both

decisions to deceive and responses to deception lies in the intended beneficiary of the lie. That is, who is the deceptive act intended to benefit—the self, or others?

Self-oriented lies are those that are intended to benefit oneself. Cheating on a test, lying on one's tax returns, and exaggerating one's knowledge or abilities are all examples of self-oriented lies. The observant reader will notice that in Chapters 2 and 3, I do not discuss self-oriented lies, but rather *selfish lies*, which I define as lies that are intended to benefit oneself, *potentially at the expense of others*. This second clause creates an important distinction between the two types of lies. With selfish lying, the deceiver knows that the lie can or will affect the target or a third party in a negative way. Self-oriented lies, on the other hand, are agnostic as to what the effect on the target or third party will be. To illustrate, imagine I tell a friend that I have not been drinking when I am in fact drunk in order to appear responsible. While he may or may not believe me, this lie would likely not have an appreciable effect on the wellbeing of my friend. Thus, this would constitute a self-oriented lie. On the other hand, if I tell this same lie in an effort to convince my friend to ride in my car with me behind the wheel, this would be a selfish lie, as it would put my friend's life in danger. The distinction between these two types of lies highlights a key component of the present theoretical framework of deception. That is, at the most basic level, lies can be distinguished solely by the intended beneficiary of the lie. The effects the lie has on the party (deceiver or target) who is not the beneficiary is indeed an important consideration that should not be ignored. Yet, there are complexities with incorporating these considerations into a theory of lying, as I will discuss later in this chapter.

Before that, however, it would be prudent to address the focus of this dissertation: prosocial lies. Just as self-oriented lies are those which are intended to benefit oneself, prosocial lies (or other-oriented lies; DePaulo et al., 1996) are intended to benefit others. More

specifically, prosocial lies are false statements that are made to intentionally, knowingly, and/or purposely mislead and benefit others (Levine & Schweitzer, 2014, 2015).[3] The meaning behind the "mislead" component of this definition should be evident from previous discussion of lying and deception; the "benefit others" component is what makes this form of lying uniquely prosocial. With prosocial lying, the primary goal of lying is to benefit another individual or group in some way. The intended benefit of prosocial lies (and self-oriented lies) can vary in tangibility. For instance, one might lie in order to procure a tangible good for another, such as money; or, one could lie in order to provide a less concrete benefit for another, such as giving someone a confidence boost. This benefit can be intended for the target of the lie. For example, one might compliment a friend on an unfortunate haircut in order to make her feel good. The benefit can also be intended for a third party. For instance, a professor might exaggerate a student's abilities on a letter of recommendation; in this case, the recipient of the letter is the target of the lie, but the student is the beneficiary of the lie. Regardless of the nature of the benefit, however, there must be some intended benefit of a lie for another individual or group for that lie to be considered prosocial.

[3] Prosocial and self-oriented lying are subsets of prosocial and self-oriented deception, respectively. I use the term "prosocial lying" because Chapters 2 and 3 describe experimental designs used to study lying specifically. However, there are ways in which people can engage in deception for the benefit of others that do not involve making false statements. Furthermore, despite the theoretical differences between lying and deception, the degree of overlap between the two constructs, as well as the popularity of lying as a form of deception, render it not unreasonable to use the terms interchangeably.

In spite of the ostensibly straightforward definition of prosocial lies, there are at least two issues that could pose problems both for construct validity of prosocial lying measures, as well as for general consistency with which the term is used in future research. The first concerns whether we define prosocial lies based on their intentions or on their consequences; the second regards how the prosocial intentions of these lies are assessed and the problem of mixed motives. Below, I address these issues to clarify the construct of prosocial lying.

Intention- and Consequence-Based Definitions of Prosocial Lies

My definition of prosocial lying takes an intention-based approach. That is, according to this definition, prosocial lies are prosocial in their intentions, and not necessarily in their consequences. While this definition is consistent with that of others who have studied prosocial lies (Levine & Schweitzer, 2014, 2015; DePaulo et al., 1996), not all deception researchers share this perspective. Erat and Gneezy (2012), for example, take a consequence-based approach, defining types of lies based on the payoffs (i.e., outcomes) for both the deceiver and the target. According to this framework, altruistic white lies help the target at the expense of the deceiver, and pareto white lies help both parties. Beyond the domain of deception, there has also been disagreement as to whether prosocial behaviors in general should be defined based on their intentions or consequences. In their seminal review of prosocial behavior, Penner et al. (2005) define prosocial behavior as "a broad category of acts that are defined by some specific segment of society and/or one's social group as generally beneficial to other people" (Penner, Dovidio, Piliavin, & Schroeder, 2005). According to this definition, the acts themselves must be beneficial to others; no mention of intentions is made. Other researchers have adopted an intention-based approach; Batson and Powell (2003) write that prosocial behaviors are actions *intended* to benefit one or more people other than oneself" (italics added). Thus, there is not widespread

agreement as to whether acts of deception (or any acts) should be considered prosocial because they have prosocial intentions or prosocial consequences.

Why all the disagreement? One glaring reason is that there is often a disconnect between intentions and consequences of behavior. That is, motivated actions do not always yield the consequences that were intended, and they can sometimes result in consequences that were not intended. Those who work hard to become rich sometimes end up poor; gifts that were meant to please go unused and forgotten; and foreign invasions aimed to foster international stability sometimes have precisely the opposite effect. Well-intended lies are no exception to this rule: People often tell lies that are meant to help others, but these lies may end up doing more harm than good. Consider the professor who writes an inordinately positive recommendation letter for a student, who uses that letter to land a prestigious job. Let's say that the professor wrote the letter with genuine concern for the wellbeing of the student. However, upon commencing the job, the student quickly realizes he is underqualified and overwhelmed, and he is soon fired. In the process, the professor himself may have damaged his reputation in the field as word spreads about the inaccuracy of his letter. For both the student and the professor, the well-intended lie backfired. I discuss backfiring of prosocial lies in detail in Chapters 2 and 3, but it bears worth repeating here: Prosocial lies do not always have prosocial consequences.

Acknowledgement of the possibility for prosocial lies to backfire sets the backdrop for differentiation between these lies and a related construct: white lies. Colloquially, white lies are sometimes used synonymously with prosocial lies; a person who compliments a friend on a bad haircut might consider herself to have told a white lie. Some researchers have echoed this perspective. Erat and Gneezy (2012), for example, define white lies as those that help others. However, white lies can also be viewed as those that involve small stakes, regardless of whether

the intention is self- or other-serving (Bok, 1978; Levine & Schweitzer, 2015). For instance, if I stole my colleague's pen and then later denied it when questioned, this could be considered a white lie in that there likely would be no profound consequences for the colleague (pens are cheap and readily available). Yet, this lie would not be told with the intent to help others in accordance with the definition of prosocial lying, but would instead be a function of my self-oriented desire to avoid culpability. Given the potential for prosocial lies and white lies to have both different intentions (other-oriented vs. self-oriented or other-oriented, respectively) and different consequences (minor or substantial vs. minor, respectively), I view these as potentially overlapping but distinct constructs.

Because prosocial lies can have harmful consequences on others, it is reasonable to question whether other-oriented intentions should be used as the defining feature of these lies at all. Defining prosocial lies based on their consequences (e.g., Erat & Gneezy, 2012; Gneezy, 2005) has several merits. First, experimental designs that operationalize prosocial lies as those which help others are likely to have greater construct validity than intention-based operationalizations. The reason for this is because in laboratory experiments, intentions are more difficult to measure accurately than consequences. In the typical consequence-based prosocial lying design, participants are given the opportunity to lie or cheat in a game that results in a tangible benefit for another person or group (e.g., a monetary payoff). Here, there is no question about whether the lie has prosocial consequences: Participant dishonesty makes the target strictly better off than honesty. In contrast, prosocial intentions are typically assessed either by self-report, or inferences based on features of the experimental design, which often amounts to prosocial consequences of the lies. Self-report of prosocial intentions are subject to the usual issues with self-report measures, which include individuals' inability to access or quantify their

motivations, beliefs, and other internal states (Nisbett & Wilson, 1977a), as well as impression management (Paulhus & Reid, 1991) and experimenter demand. A well-designed experiment can increase researchers' confidence that individuals made a decision out of prosocial concerns without the use of self-report. For example, if one lies to increase another's payoff at a cost to one's own payoff, it may be assumed that the deceiver intended to benefit the target. Yet, even this type of design is not free from experimenter demand, as people may lie not because of their prosocial preferences, but in order to be seen in a positive light by the experimenter. Thus, defining prosocial lies based on prosocial consequences (versus intentions) alleviates concerns about construct validity stemming from measurement error and potential confounds.

Another advantage of a consequence-based approach is that it lends well to uncovering how incentives influence lying for the benefit of others (as well as oneself). With economic games, researchers can directly link decisions to lie with specific outcomes for oneself and others. For example, in the Sender-Receiver Game (Erat & Gneezy, 2012; Gneezy, 2005; Gneezy, Rockenback, & Serra-Garcia, 2013; Zhong, 2011), participants choose whether to send an honest or dishonest message to a partner, which then results in both players being paid amounts that were known to the sender prior to the decision. As such, one can quantify how expected outcomes for the self and others affect willingness to lie. Researchers have shown with this type of design that a substantial number of people will lie when it costs them a little and helps others a lot. Furthermore, some people exhibit lying aversion, such that they will not lie even when it helps oneself and a partner equally (Erat & Gneezy, 2012). By clearly tying decisions to lie with outcomes, researchers can strengthen the internal validity of experiments on lying by reducing the potential for confounding variables. In addition, this approach can ideally

allow us to extrapolate results from games to better understand how expected outcomes influence decisions to lie in real world contexts.

In spite of the merits of defining lies based on their consequences, however, there are also advantages of intention-based definitions. In my view, these advantages outweigh the drawbacks, which is why I take an intention-based approach to prosocial lying. Why should we classify prosocial lies based on their intentions and not their consequences? Consider the following reasons:

Consequences are uncertain. The ultimate consequences of prosocial lies are not always apparent or even knowable to deceivers, targets, and/or observers. Let's return to the example of the professor who writes an overly positive recommendation letter for a student, who then successfully secures a job for which he is underqualified. This time, however, imagine that the student is not fired from the new position. Although the student is happy with his salary, he is overwhelmed by the work and constantly under stress. In this example, to what extent did the professor's lying about the student's credentials have prosocial consequences? As far as the student is concerned, it depends how the utility gained by the money he is being paid compares with utility lost due to the stress of the work. This may be a challenging exercise in self-knowledge for the student. It may also be difficult if not impossible for the professor or other outside observers to determine whether writing the deceptive recommendation letter made the student better off.

Of course, there are situations where one a lie has clear and immediate positive consequences for others. For example, imagine a weight-conscious friend asks you how she looks in an outfit. Believing that the truth will hurt her feelings, you tell her that she looks wonderful, and the positive effects of this statement are visible on her face. However, even in

this example, and in other cases in which the consequences of a lie are apparent, there may be unintended and unforeseeable downstream consequences. Perhaps as a result of receiving positive reinforcement from you, your friend wears this outfit on a regular basis, and later gets dirty looks or ridicule from others, which completely eradicates all emotional benefits she received from you. Perhaps your reinforcement subconsciously signals to her that she doesn't need to be so vigilant about her weight, which then leaders her to exercise less. The point here is that even in cases where a lie (or any behavior) clearly causes a result (where causal attribution is being made due to temporal closeness of the cause and effect, and lack of plausible alternatives; Cheng & Novick, 1992; Wells & Gavanski, 1989), there is uncertainty surrounding the downstream effects of the behavior. This uncertainty makes it more difficult to study the consequences of lying.

Another problem to consider with this example is that even if there is a net increase in utility for the student, the professor's lie arguably had a negative effect on others—that is, the hiring organization, and other job candidates. By deceiving the organization, the professor may have prevented a more highly qualified candidate from being hired. This is unfair to the other candidates, and detrimental to the organization, which may have functioned more efficiently with a different hire. As such, consequence-based definitions of prosocial lies can be problematic because they raise the question of, "prosocial for whom?" A single lie can have positive consequences for one party and negative consequences for another party.

A third issue related to the uncertainty of consequences concerns the problem of causality: To what extent did the lie cause the consequences in question? To return to the aforementioned example once more, imagine that the professor's letter was used as part of a larger application package, which included a CV, transcripts, essays, and other letters of

recommendation. Would the student have gotten the job had the professor not written him a letter? It may be difficult or impossible to assess causality here since the letter in question made up only one component of the student's application. The letter may have been the one aspect of the application to push the student over the edge into the "hirable" pile; it may have been an application piece that matched up with other candidates' letters, whereby another part of the application was the dealmaker; and it may also have been disregarded by the committee, who put little weight on the letters aside from when they contain glaring red flags. There are many other circumstances in which it difficult to assess causal relationships between events, such as when there is a large temporal distance between events, or when there are many possible causes. Using economic games such as those described earlier is one way of circumventing this problem; as mentioned, with economic games there is a direct link between decisions to lie or tell the truth and the consequences of those decisions. However, the real world is messier than this, and determining causal effects of prosocial lies outside the lab is a more challenging task. Given that identifying causal effects of prosocial lies may be problematic, a focus on intentions rather than consequences of these lies is one solution that researchers can implement.

Intentions are interesting and important. A second reason why I employ an intention-based definition of prosocial lying is that intentions of these lies is an interesting and important area of study in itself. One practical guideline for determining if a theory is interesting is asking whether it violates widely but weakly held assumptions (Davis, 1971). The idea that "lying is wrong" meets this guideline: In general, people say they value honesty (Graham et al., 2013), yet people frequently lie (DePaulo et al., 1996). Though people do lie for selfish reasons (which likely is a primary reason why people think that lying is wrong; Buller & Burgoon, 1996), an easing of the moral opposition to dishonesty sometimes comes when lying is seen as helpful to

others (Levine & Schweitzer, 2014, 2015). How do individuals navigate the tension between the moral values of honesty and kindness (i.e., harm/care; Graham, Haidt, & Nosek, 2009; Graham et al., 2011)? This question lies at the core of understanding what drives people to tell prosocial lies, as well as how deceivers' intentions influence responses to these lies. One aspect of the interestingness of this question stems from the lack of an easier answer for navigating this moral dilemma—saying that "lying is wrong" doesn't hold as a blanket statement here, at least for most people, and there are many considerations to take when determining whether one should tell a prosocial lie as well as how one should view these lies.

Another component of the interestingness stems from the implications themselves: Prosocial lies can have profound effects on people's lives. While I discuss the implications of prosocial lies in Chapters 2 and 3, I will say here that focusing on the intentions of these lies is critical for predicting and understanding what the consequences will be. This is important because, given that prosocial lies can backfire, we want to know when we might be the target of a prosocial lie; when we might be likely to lie prosocially to others; and how people respond to these lies in general. Through this knowledge, we can hopefully improve interpersonal interactions by having a more thorough understanding of when prosocial lying is and is not seen as appropriate and helpful to others.

To be clear, I am not saying that we should not study the consequences of lying. In fact, examining the consequences (or what we believe the consequences to be) of prosocial lies is essential to understanding how people respond to these lies, and whether telling these lies is a good social strategy. In Chapter 3, I take a hybrid approach by breaking down prosocial lying into two subsets: paternalistic lies (those that require the deceiver to make assumptions about the target's best interests) and unequivocal prosocial lies (those that yield unequivocal benefits for

the target). As I will discuss in that Chapter, this is a critical distinction in determining how people respond to prosocial lies. However, as far as the superordinate form of prosocial lying goes, defining a lie to be prosocial based on its intentions rather than its consequences is preferable because (a) it may be difficult or impossible to know what consequences resulted from well-intended lies and (b) the motivation behind prosocial lying is an interesting and important phenomenon of study in itself.

Measuring Prosocial Lying and Prosocial Intentions

Once we agree that an intention-based definition of prosocial lying is acceptable and that this construct is worthy of study, the question then becomes: How can we measure prosocial lying? More specifically, how do we assess whether a deceiver has prosocial motivation? Furthermore, how do we account for mixed motivations (i.e., both self-oriented and prosocial intentions)?

In its relatively short history of study, prosocial lying has been examined primarily using three approaches:[4]

[4] Note that these approaches pertain only to studying the telling of prosocial lies; the investigation of responses to prosocial lies requires different designs, albeit with some overlap between approaches. In general, studying responses to prosocial lies first necessitates observation of prosocial lying, either imaginary (via hypothetical vignettes) or real (through any of the three design-types described above, or witnessing these lies in interpersonal interactions in person, over video, or in written form). Then, responses are assessed via self-report (e.g., moral judgment scales, written free-responses) or behavioral (e.g., punishment) measures. It is also possible to investigate responses to prosocial lies in interpersonal contexts, where both the

Self-report. One way that prosocial lying has been studied is through self-report. Self-reporting of prosocial lying has been measured in a variety of ways. Earlier studies have employed diary studies, where participants are asked to record details about each lie they tell on a daily basis for a period of time. Some of these studies have asked participants to report specifically on "white lies" (Camden, Motley, & Wilson, 1984), while others did not specify the type of lie, and later used coders to determine whether the lie was self- or other-oriented (DePaulo et al., 1996; DePaulo & Kashy, 1996; Kashy & DePaulo, 1998). Researchers have also asked participants to recall and recount the prosocial lies they have told, as well as to describe these lies in interviews (Turner, Edgley, & Olmstead, 1975; Hample, 1980).

This body of work takes an exploratory, bottom-up approach to studying prosocial lies, whereby we learn about prosocial lies straight from the mouths of those who tell them: to whom people tell prosocial lies, what they lie about, the reported easons for lying, etc. This strategy has the benefit of helping us identify what prosocial lying looks like in the real world, as well as the frequency with which it occurs. It can also and gives us a sense of how lay people think about these lies. Yet, there are limitations to this approach as well. First, people's ability to accurately recall instances of lying and the reasons behind them may be limited. Social desirability likely biases accounts of lying such that people overestimate the degree to which other-oriented considerations motivate deceptive behavior.[5] This approach also limits our ability to make causal

---

deceiver and target are actively engaging face-to-face and in real time. However, as described in the main text, this approach has yet to be implemented by researchers.

[5] Although, people do also report instances of self-oriented lying, and they report telling these lies more often than prosocial lies (DePaulo et al., 1996).

inferences about drivers of prosocial lying, as it is not experimental. Despite these issues, however, these bottom-up self-report studies have laid the foundation for our understanding of prosocial lying at the phenomenological level.

A second way in which prosocial lies are studied with self-report is by first observing lies told in real time in the laboratory, and then asking participants about the reasons behind lying, either via quantitative measures or qualitative free-response narratives. If participants are given the opportunity to tell a lie that has real or perceived benefits for others, and then report that their decision was made because of their motivation to help others, this can raise our confidence that the lie was prosocially motivated. Likewise, self-report can be used to rule out that factors besides the concern for others' wellbeing motivated the decision to lie. Impression management and social desirability can be problematic here, but those concerns can be mitigated by carefully designed experiments that reduce the plausibility of these and other third-variable explanations (e.g., structuring the design so that participants believe that the experimenter cannot observe deceptive behavior).

Of course, an important question is how to design such an experiment so that prosocial lying can be observed in the laboratory. This can be accomplished through two other methods for studying these lies: economic games and quasi-interpersonal interactions.

Economic "games". Another way prosocial lying has been measured is through economic games. I use the term "game" loosely here, as not all of these designs contain multiple players. However, all of these designs do contain some variation of the following elements: (a) participants take part in a game or other procedure that has a set of rules or instructions; (b) those rules can be broken through the use of deception, *or* engaging in deception is accepted under the rules; (c) the use of deception directly results in (or increases the chances of) payoffs to other

individuals or groups. In these designs, participants have the opportunity to deceive either the experimenter, another individual (usually another participant who is either real or fictional), or group of people. The target is often another participant(s) in the experiment (real or fictional), or a charity. In addition, the potential target of deception need not be the beneficiary of the lie, though this is sometimes the case. Finally, the payoff need not be monetary; participants can lie so that others receive benefits such as non-monetary goods or exemption from a tedious task.

Researchers have employed several variants of these games. The aforementioned Sender-Receiver Game is one such example. In this game, deception is baked into the design; participants can send either a false or true statement to a partner (either real or imaginary) about something arbitrary, such as the outcome of a die roll or coin flip, and the false statement results in the partner earning more money than the true statement. Another popular design is the matrix task, whereby participants are asked to solve a series of mathematical puzzles, and are given the opportunity to covertly (or so they believe) over-report the number of matrices solved (Mazar, Amir, & Ariely, 2008). Unbeknownst to participants, researchers devise clever ways determine how many matrices were actually solved, such as by rummaging through the trash to find participants' worksheets. By tying reported performance on the task to the payment of other participants and measuring the difference between actual and reported performance, researchers can obtain a measure of prosocial lying (Gino, Ayal, & Ariely, 2013; Wiltermuth, 2011). There are several variations on this task, which include over-reporting the number of anagrams solved, the number of times a coin or die lands on a particular face, or the number of times a series of dots falls on one side of a diagonal on a screen (Bryan, Adams, & Monin, 2013; Gino, Norton, & Ariely, 2010; Gino & Pierce, 2009). Other variants are likely to emerge. The key commonalities of all these designs, however, are that participants can lie, cheat, or otherwise misrepresent the

true state of the world to the experimenter; experimenters can observe or infer dishonesty, either through measurement (e.g., reported minus actual performance) or probabilistic calculation (e.g., reporting that a coin lands on heads a number of times that would be exceedingly rare based on expected probability); and dishonesty results in benefits for others.

One might ask: don't these games assume a consequence-based definition of prosocial lying? Indeed, they do. However, as mentioned earlier, consequence-based designs can be used to infer prosocial motivation if dishonesty is tied only to the payments of others rather than the self. In addition, one can implement self-report measures as a follow-up to these designs to gain greater confidence and precision about the driving mechanism for dishonesty (e.g., emotions; Gino & Pierce, 2009).

Economic games to measure prosocial lies are subject to the same merits (construct validity, use of incentives) and drawbacks (lack of external validity, potential for demand) of designs to measure consequence-based prosocial lying as discussed earlier. They also have the added benefit over diary and recall-based methods of allowing for causal inference. That is, by varying features of the game between-subjects (e.g., incentives), or implementing other pre-deception decision experimental treatments (e.g., emotion manipulation), one can acquire evidence for causal drivers of prosocial lying. A final advantage of these games is that they allow for two different operationalizations of prosocial lying: the rate of prosocial lying (i.e., whether participants lie at all) and the magnitude of prosocial lying (i.e., the strength or amount of dishonest behavior). While two measures of prosocial lying per experiment may be better than one from the perspective of a researcher trying to publish, consideration should be given to the theoretical relevance of each measure, as measures drawn from different designs may have different real world parallels (or lack thereof).

Quasi-interpersonal interactions. Given the discrepancy between the nature of economic games in which prosocial lies are measured and the real world contexts in which these lies are told, researchers have begun to develop experimental designs that more closely approximate the latter. The result is designs that include quasi-interpersonal interactions.

In these designs, participants are told that they will interact or communicate with another individual over a digital medium. This communication may or may not occur in real time, and the other individual may or may not be fictional. Participants have the opportunity to tell prosocial lies to their partner, which can be verified as deceptive by the experimenter. One example of this design is implemented in Studies 1 and 2 of Chapter 2. While the procedure is described in detail in Chapter 2, the general idea is that participants first provide private (i.e., to the experimenter only) evaluations of their partner's performance on a task. Afterwards, they evaluate their partner's performance again on the same measures, except now with the instructions that their partner will view these evaluations. Prosocial lying can be operationalized as the difference score of shared evaluations minus private evaluations; motivations and alternative explanations can be assessed with self-report measures after evaluations are made.

As this type of design has been implemented relatively recently in the study of prosocial lying (Jampol & Zayas, 2018), researchers are likely to develop variations on this type of design to improve construct and external validity. However, the core feature of quasi-interpersonal designs is that participants have the opportunity to tell a prosocial lie to another individual, but participants and their counterparts do not interact face-to-face. This type of design has some similarities with economic games to study prosocial lies. Both classes of designs involve behavioral measures of prosocial lying in the laboratory or online. Like with economic games, researchers can use quasi-interpersonal methods to test causal predictors of prosocial lying by

embedding experimental treatments within the design. An important difference between these types of designs, however, is that that in quasi-interpersonal interactions, participants tell prosocial lies about topics that could more realistically occur in real interpersonal interactions (e.g., performance feedback vs. the number of word puzzles solved). Thus, while the medium of communication may not resemble that of normal interactions, participants are lying to other people about things that they might lie about in the real world. Another difference between economic games and quasi-interpersonal designs (as they've been implemented thus far) is that the latter take a more intention-based approach to prosocial lying. That is, in quasi-interpersonal interaction methods, researchers have examined the factors that influence prosocial lying, without necessarily exploring the downstream effects lying has on the target. Although, it is possible to implement consequences into these designs, either by manipulating the effects their deception decision has on their partner (which can be specified before the decision is made or not), or (in cases where there is a real interaction partner) allow an actual response from the partner. As designs like these are developed in the future, prosocial lying researchers will build an experimental toolkit that facilitates a more precise glimpse into the causes and consequences of prosocial lies.

The use of quasi-interpersonal designs for the study of prosocial lies begs the question: Shouldn't we study prosocial lying in the context of real interpersonal interactions? I believe that we should, and studies with this type of design are likely to appear in the future. To date, however, prosocial lying has not been studied in real-time social interactions. Part of the reason for this likely has to do with logistical (e.g., scheduling participant dyads, recruiting and training behavioral coders) and analytical (e.g., accounting for interdependence, actor, partner, and dyad effects) difficulties of running this type of experiment. Yet, these difficulties are not

insurmountable, and dyadic interaction studies can enrich our understanding of prosocial lying. By studying these lies in real interaction contexts, researchers can gain more confidence in the external validity of experimental treatments to influence prosocial lying, as any prosocial lying will be occurring face-to-face with another human being. They can also answer questions such as: What communicative strategies (e.g., lying, omission, paltering, etc.) do people use to deceive others for their perceived benefit? Do prosocial lies elicit the same nonverbal "tells" as self-oriented lies, and are these lies more difficult to detect than self-oriented lies? How does knowing that one has been told a prosocial lie in conversation influence affective responses and perceptions of the deceiver? How effective are face-to-face apologies or justifications for prosocial lying? And how do decisions to lie and responses to these lies differ by type of relationship? These are all interesting research questions that dyadic interaction studies can help to answer.

How Prosocial is Prosocial Lying?

The notion that people sometimes act out of genuine consideration for the wellbeing of others is a relatively uncontroversial idea. However, what is controversial is whether people ever behave *solely* out of consideration for others. Whether any act can be purely altruistic (i.e., of benefit to others and either no benefit or harm to oneself) has been debated extensively by psychologists (Batson, 1987; Batson & Shaw, 1991; Campbell, 1975; Cialdini et al., 1997; Rushton, 1989) and philosophers (Bentham, 1789/1879; Comte, 1851/1875; Hume, 1740/1896; Nagel, 1970). These arguments are beyond the scope of this paper. However, this debate does bear relevance to understanding the construct of prosocial lying because it raises the question of, to what extent can self-interest contribute to one's motivation to tell a prosocial lie before that lie can no longer be considered prosocial?

One perspective that sheds light on this issue stems from theory on impure altruism (Andreoni, 1989; 1990). According to this theory, people do want to help others, but they do so in part because of the "warm glow" or positive feelings that come with doing good for others. It is helpful to view prosocial lying through the lens of impure altruism: Individuals lie with the intention to benefit others, but they also may derive affective benefits themselves by doing so. In my view, the existence of these benefits to oneself do not negate the prosocial nature of prosocial lying, so long as the lie is not told primarily for the purpose of benefitting oneself. Thus, prosocial lying is just as "impure" as any other prosocial act.

However, the theory impure altruism alone does not address when a lie should or not be considered prosocial given the presence of self interest. People sometimes tell lies that are intended to help others, but that also result in advantages for oneself apart from warm glow. For instance, an individual might lie to his/her significant other about an infidelity to protect that person's feelings, but also so that s/he may continue to reap benefits of the relationship (e.g., emotional support, financial security, sex). In some cases, these mixed motives may be obvious. But there are other situations where it is ambiguous whether a lie is motivated by prosocial or self-oriented considerations (or both), from the perspective of targets, observers, and perhaps even deceivers. In the above scenario, for example, the deceiver may be unaware of the extent to which his/her own self-interest played into the decision to deceive his/her partner. Apart from the problem this ambiguity of motives poses to the measurement of prosocial lying (as discussed in the "Measuring Prosocial Lying and Prosocial Intentions" section), it also highlights the difficulty of defining prosocial lies based on intentions given that self-interest can cloud what on the surface appear to be prosocial acts.

While there is no cure-all solution to this problem, one potential approach is to view prosocial lies as that those are at least in part intended to help others, regardless of the degree to which self-oriented motivation influences the behavior. This perspective is consonant with that of Erat and Gneezy (2012), who differentiate between altruistic white lies (those that require a sacrifice on the part of the deceiver to help the target), and pareto white lies (those that help both the deceiver and the target). The analogous perspective in the current framework would be that prosocial lies can be both altruistic or pareto in their intentions—that is, a lie is prosocial so long as it is intended to help others in some way, regardless of the intended benefit for oneself. Viewing prosocial lies in this way solves the issue of mixed motives simply by allowing mixed motives to exist within the construct. However, one problem with this view is that it may come at the expense of face validity; a lie that is expected to yield a very large payoff for oneself and a very small payoff for another can hardly be considered prosocial in the normative sense of the word. Although lies of this nature may be a worthy area of study, in some contexts they may be sufficiently different from lies that are primarily intended to help others such that lumping them together comes at a loss of explanatory power.

Another way to approach the issue is to consider prosocial lies to be those that are *primarily* intended to benefit others. In other words, self-interest is a tolerated component of prosocial lying so long as prosocial considerations constitute the majority of the reason why the lie was told. Of course, the question then becomes, how can we tell if a lie is primarily prosocial? It is difficult to accurately measure the proportion of an individual's self-oriented versus other-oriented motivation. However, there are strategies that researchers can use to gain confidence that self-oriented motivation did not drive decisions to tell what is ostensibly a prosocial lie. First, one must carefully design experiments to assess the role of self-interest. Some ways this

can be done were already mentioned in the discussion of methods to measure prosocial lies; for example, economic games used to measure prosocial lying can also be structured so that lying helps others but comes at a cost to oneself. If costly helping via lying is observed, this suggests the presence of prosocial rather than self-oriented motivation. Ensuring that decisions to prosocial lie are perceived as private (i.e., not observable to the experimenter) is another way researchers can attempt to rule out the influence of self-interest. If lying is clearly beneficial to others, thinking that one's decision about whether to lie is observable may increase reputational or self-presentation concerns. Thus, people might lie more to appear prosocial rather than to be prosocial. Absent the perception that others will know whether or not one lies, this concern is mitigated.[6] In addition, researchers can craft experimental treatments where it is possible to escape the decision to lie, or otherwise disguise the relationship between their decision and outcomes for the beneficiary of the potential lie (e.g., Dana, Weber, & Kuang, 2007). If more people choose to escape or make no decision rather than to tell a prosocial lie, this suggests that decisions to lie in treatments where no escape was possible were at least in part driven by feelings of obligation or other non-prosocial concerns. Finally, self-report measures can supplement features of experimental designs to measure self-oriented motivation. Inclusion of these measures along with self-report measures of prosocial motivation in multiple mediation models can help to determine the processes participants believed factored into their decision

[6] Experimentally manipulating whether the decision to tell a prosocial lie is made publicly or privately can help to assess people's beliefs about how others view these lies—an interesting and underdeveloped area of study. If more prosocial lying is observed in public (vs. private), this suggests that people think others view prosocial lying favorably.

making. While it may be difficult if not impossible to rule out with complete certainty whether decisions to prosocial lies were made absent of self-interest, by implementing several of these strategies in conjunction and obtaining convergent results with multiple methodologies, researchers can increase their confidence that prosocial lies were indeed prosocially motivated.

Tables

Table 1.1: Definition of terms with examples.

Construct	Definition	Example(s)	Notes
Deception	The transmission of information to intentionally, knowingly, and/or purposely mislead others.	Lying, deceptive omission, paltering, cheating, disguises, encrypted language.	Superordinate term for all intentionally misleading behaviors.
Lies	False statements made to intentionally, knowingly, and/or purposely mislead others.	Telling a boss that work has been completed when it hasn't to avoid punishment.	Sometimes used interchangeably with deception, though, lying is a subset of deception. Not all acts of deception involve lying (e.g., deceptive omission, paltering, etc.).
Prosocial Behavior	A broad range of actions intended to benefit one or more people other than oneself (Batson & Powell, 2003).	Charitable giving, volunteering, helping, comforting, sharing, cooperating, caring for offspring.	Other definitions have focused on consequences. For example, Penner et al. (2005) define prosocial behavior as acts that are in some way beneficial to others.
Prosocial Lies	False statements made to intentionally, knowingly, and/or purposely mislead and benefit others.	Telling someone they performed well when you believe they performed poorly (to protect their feelings); A doctor telling a patient that they are likely to live longer than than is truly believed (to foster hope).	In this framework, the "prosocial" in prosocial lying refers to intentions, not outcomes. A prosocial lie might not be in the best interests of the target, in the eyes of the target or observers. Other forms of deception (i.e., besides lying) can be prosocial. Synonymous with what DePaulo et al. (1996) call other-oriented lies ("Lies told to protect or enhance other persons psychologically or to advantage or protect the interests of others").

Table 1.1, Continued: Definition of terms with examples.

White Lies	Relatively harmless lies that have little or no consequences for the recipient or others (Argo & Shiv, 2011; Camden, Motley, & Wilson, 1984). According to this definition, lies can be self-oriented or other-oriented.	Telling a colleague that you did not steal one of her tissues when in fact you did.	According to this definition, a white lie could be a prosocial lie, but is not necessarily. Prosocial lies may or may not be inconsequential. White lies have also been defined as those help others (Erat & Gneezy, 2012). In my view, this definition is closer to *unequivocal prosocial lies*, or prosocial lies that have positive consequences for others (See Chapter 3).
Self-oriented lies	False statements made with the intent to mislead others and benefit oneself (DePaulo et al., 1996).	Misrepresenting one's income on tax returns to avoid paying taxes.	Self-oriented lies are a superset of selfish lies, which also are intended to benefit oneself, but which come with a potential cost to others. Self-oriented lies are agnostic to the benefit or cost to others.

References

Andreoni, J. (1989). Giving with impure altruism: Applications to charity and Ricardian equivalence. *Journal of Political Economy*, *97*(6), 1447-1458.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, *100*(401), 464-477.

Batson, C. D. (1987). Prosocial motivation: Is it ever truly altruistic?. In *Advances in Experimental Social Psychology* (Vol. 20, pp. 65-122). Academic Press.

Batson, C. D., & Powell, A. A. (2003). Altruism and prosocial behavior. *Handbook of Psychology*.

Batson, C. D., & Shaw, L. L. (1991). Evidence for altruism: Toward a pluralism of prosocial motives. *Psychological Inquiry*, *2*(2), 107-122.

Bok, S. (1978). *Lying: Moral Choices in Public and Private Life.* New York: Pantheon.

Boles, T. L., Croson, R. T., & Murnighan, J. K. (2000). Deception and retribution in repeated ultimatum bargaining. *Organizational Behavior and Human Decision Processes*, *83*(2), 235-259.

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication Theory*, *6*(3), 203-242.

Bryan, C. J., Adams, G. S., & Monin, B. (2013). When cheating would make you a cheater: Implicating the self prevents unethical behavior. *Journal of Experimental Psychology: General*, *142*(4), 1001.

Campbell, D. T. (1975). On the conflicts between biological and social evolution and between psychology and moral tradition. *American Psychologist*, *30*(12), 1103.

Chance, Z., Norton, M. I., Gino, F., & Ariely, D. (2011). Temporal view of the costs and benefits of self-deception. *Proceedings of the National Academy of Sciences*, *108*(Supplement 3), 15655- 15659.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*(2), 365.

Cialdini, R. B., Brown, S. L., Lewis, B. P., Luce, C., & Neuberg, S. L. (1997). Reinterpreting the empathy–altruism relationship: When one into one equals oneness. *Journal of Personality and Social Psychology*, *73*(3), 481.

Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, *1*(1), 49-98.

Ekman, P., & Friesen, W. V. (1974). Detecting deception from the body or face. *Journal of Personality and Social Psychology*, *29*(3), 288.

Ekman, P., & O'sullivan, M. (1991). Who can catch a liar?. *American Psychologist*, *46*(9), 913.

Erat, S., & Gneezy, U. (2012). White lies. *Management Science*, *58*(4), 723-733.

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*(1), 67-80.

Davis, M. S. (1971). That's interesting! Towards a phenomenology of sociology and a sociology of phenomenology. *Philosophy of the Social Sciences*, *1*(2), 309-344.

DePaulo, B. M., & Kashy, D. A. (1998). Everyday lies in close and casual relationships. *Journal of Personality and Social Psychology*, *74*(1), 63.

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology*, *70*(5), 979.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *129*(1), 74.

Ekman, P., & Friesen, W. V. (1974). Detecting deception from the body or face. *Journal of Personality and Social Psychology*, *29*(3), 288.

Gino, F., Ayal, S., & Ariely, D. (2013). Self-serving altruism? The lure of unethical actions that benefit others. *Journal of Economic Behavior & Organization*, *93*, 285-292.

Gino, F., Norton, M. I., & Ariely, D. (2010). The counterfeit self: The deceptive costs of faking it. *Psychological Science*, *21*(5), 712-720.

Gino, F., & Pierce, L. (2009). The abundance effect: Unethical behavior in the presence of wealth. *Organizational Behavior and Human Decision Processes*, *109*(2), 142-155.

Gino, F., & Shea, C. (2012). Deception in Negotiations. *The Oxford Handbook of Economic Conflict Resolution*, 47.

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, *95*(1), 384-394.

Gneezy, U., Rockenbach, B., & Serra-Garcia, M. (2013). Measuring lying aversion. *Journal of Economic Behavior & Organization*, *93*, 293-300.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in Experimental Social Psychology* (Vol. 47, pp. 55-130). Academic Press.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366.

Hample, D. (1980). Purposes and effects of lying. *Southern Speech Communication Journal*, *46*(1), 33-47.

Immanuel, K. (1785). Groundwork for the Metaphysics of Morals.

Kashy, D. A., & DePaulo, B. M. (1996). Who lies?. *Journal of Personality and Social Psychology*, *70*(5), 1037.

L. Knapp, M., & Comaden, M. E. (1979). Telling It like It Isn't: A Review of Theory and Research on Deceptive Communications. *Human Communication Research*, *5*(3), 270-285.

Jampol, L., & Zayas, V. (2017). Gendered White Lies: Performance Feedback is Upwardly Distorted to Women. *Working Paper*.

Kunst-Wilson, W. R., & Zajonc, R. B. (1980). Affective discrimination of stimuli that cannot be recognized. *Science*, *207*(4430), 557-558.

Levine, T. R. (2014). Truth-Default Theory (TDT) A Theory of Human Deception and Deception Detection. *Journal of Language and Social Psychology*, *33*(4), 378-392.

Levine, E. E., & Schweitzer, M. E. (2014). Are liars ethical? On the tension between benevolence and honesty. *Journal of Experimental Social Psychology*, *53*, 107-117.

Levine, E. E., & Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, *126*, 88-106.

Lupoli, M. J., Jampol, L., & Oveis, C. (2017). Lying because we care: Compassion increases prosocial lying. *Journal of Experimental Psychology: General*, *146*(7), 1026.

Lupoli, M. J., Levine, E. E., & Greenberg, A. E. (2018). Paternalistic lies. *Organizational Behavior and Human Decision Processes*, *146*, 31-50.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633-644.

McCornack, S. A. (1992). Information manipulation theory. *Communications Monographs*, *59*(1), 1-16.

McCornack, S. A., & Levine, T. R. (1990). When lies are uncovered: Emotional and relational outcomes of discovered deception. *Communications Monographs*, *57*(2), 119-138.

McCornack, S. A., Levine, T. R., Solowczuk, K. A., Torres, H. I., & Campbell, D. M. (1992). When the alteration of information is viewed as deception: An empirical test of information manipulation theory. *Communications Monographs*, *59*(1), 17-29.

McCornack, S. A., Morrison, K., Paik, J. E., Wisner, A. M., & Zhu, X. (2014). Information manipulation theory 2: a propositional theory of deceptive discourse production. *Journal of Language and Social Psychology*, *33*(4), 348-377.

Murphy, S. T., & Zajonc, R. B. (1993). Affect, cognition, and awareness: affective priming with optimal and suboptimal stimulus exposures. *Journal of Personality and Social Psychology*, *64*(5), 723.

Nisbett, R. E., & Wilson, T. D. (1977b). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, *35*(4), 250.

Nisbett, R. E., & Wilson, T. D. (1977a). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231.

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology*, *60*(2), 307.

Penner, L. A., Dovidio, J. F., Piliavin, J. A., & Schroeder, D. A. (2005). Prosocial behavior: Multilevel perspectives. *Annual Review of Psychology.*, *56*, 365-392.

Porter, S., & Ten Brinke, L. (2008). Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions. *Psychological Science*, *19*(5), 508-514.

Rogers, T., Zeckhauser, R., Gino, F., Norton, M. I., & Schweitzer, M. E. (2017). Artful paltering: The risks and rewards of using truthful statements to mislead others. *Journal of Personality and Social Psychology*, *112*(3), 456.

Rushton, J. P. (1989). Genetic similarity, human altruism, and group selection. *Behavioral and Brain Sciences*, *12*(3), 503-518.

Schweitzer, M. E., Hershey, J. C., & Bradlow, E. T. (2006). Promises and lies: Restoring violated trust. *Organizational Behavior and Human Decision Processes*, *101*(1), 1-19.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological Bulletin*, *103*(2), 193-210.

Taylor, S. E., & Brown, J. D. (1994). Positive illusions and well-being revisited: separating fact from fiction. *Psychological Bulletin*, *116*(1.21-27).

Taylor, S. E., Kemeny, M. E., Reed, G. M., Bower, J. E., & Gruenewald, T. L. (2000). Psychological resources, positive illusions, and health. *American Psychologist*, *55*(1), 99.

ten Brinke, L., Stimson, D., & Carney, D. R. (2014). Some evidence for unconscious lie detection. *Psychological Science*, *25*(5), 1098-1105.

Turner, R. E., Edgley, C., & Olmstead, G. (1975). Information control in conversations: Honesty is not always the best policy. *Kansas Journal of Sociology*, 69-89.

Tyler, J. M., Feldman, R. S., & Reichert, A. (2006). The price of deceptive behavior: Disliking and lying to people who lie to us. *Journal of Experimental Social Psychology*, *42*(1), 69-77.

Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology*, *56*(2), 161.

Wiltermuth, S. S. (2011). Cheating more when the spoils are split. *Organizational Behavior and Human Decision Processes*, *115*(2), 157-168.

Yeung, L. N., Levine, T. R., & Nishiyama, K. (1999). Information manipulation theory and perceptions of deception in Hong Kong. *Communication Reports*, *12*(1), 1-11.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, *35*(2), 151.

Zhong, C. B. (2011). The ethical dangers of deliberative decision making. *Administrative Science Quarterly*, *56*(1), 1-25.

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and Nonverbal Communication of Deception1. In *Advances in Experimental Social Psychology* (Vol. 14, pp. 1-59). Academic Press.

Zuckerman, M., & Driver, R. E. (1985). Telling lies: Verbal and nonverbal correlates of deception. *Multichannel Integrations of Nonverbal Behavior*, 129-147.

Chapter 2

Lying Because We Care:

Compassion Increases Prosocial Lying

Matthew J. Lupoli[1], Lily Jampol[2,3], and Christopher Oveis[1]

University of California, San Diego[1]

Queen Mary, University of London[2]

London Business School[3]

Author note: The findings presented in this paper have been presented at the following conferences: *17th Annual Meeting of the Society for Personality and Social Psychology* (2016); *Annual Meeting of the Society for Affective Science, Positive Emotions Preconference* (2015); *16th Annual Meeting of the Society for Personality and Social Psychology* (2015); *35th Annual Society for Judgment and Decision Making Conference* (2014). The contents of the paper have not been posted online or shared with the media in any way.

Correspondence concerning this article should be addressed to Matthew J. Lupoli, Rady School of Management, University of California, San Diego, Wells Fargo Hall #4W124, La Jolla, CA 92093-0553 or to Christopher Oveis, Rady School of Management, University of California, San Diego, Wells Fargo Hall #4W120, La Jolla, CA 92093-0553. E-mail: matthew.lupoli@rady.ucsd.edu or coveis@ucsd.edu

Abstract

Prosocial lies, or lies intended to benefit others, are ubiquitous behaviors that have important social and economic consequences. Though emotions play a central role in many forms of prosocial behavior, no work has investigated how emotions influence behavior when one has the opportunity to tell a prosocial lie—a situation that presents a conflict between two prosocial ethics: lying to prevent harm to another, and honesty, which might also provide benefits to the target of the lie. Here, we examine whether the emotion of compassion influences prosocial lying, and find that compassion causally increases and positively predicts prosocial lying. In Studies 1 and 2, participants evaluated a poorly written essay and provided feedback to the essay writer. Experimentally induced compassion felt towards the essay writer (Study 1) and individual differences in trait compassion (Study 2) were positively associated with inflated feedback to the essay writer. In both of these studies, the relationship between compassion and prosocial lying was partially mediated by an enhanced desire to prevent emotional harm. In Study 3, we found moderation such that experimentally induced compassion increased lies that resulted in financial gains for a charity, but not lies that produced financial gains for the self. This research illuminates the emotional underpinnings of the common yet morally complex behavior of prosocial lying, and builds on work highlighting the potentially harmful effects of compassion—an emotion typically seen as socially beneficial.

Preface

When I first began this project, I started with a simple question: What leads people to tell prosocial lies? Rather than attempt to identify all dispositional and situational characteristics that predict these lies, I took a "least common denominator" approach—that is, to determine the variable(s) that parsimoniously explain prosocial lying across people and contexts.

On one hand, the answer seems obvious: Prosocial lies are intended to help others, so people tell these lies when they want to help. However, this alone is not a satisfactory answer given the limited ability of cognitive assessments of need to spur people to action. Perceptions of need can motivate prosocial behavior (Bekkers & Wiepking, 2011), but there are many cases in which information about need fails to move people to help those who need it most (Kogut & Ritov, 2005a; 2005b, Small & Loewenstein, 2003; 2005). Sadly, pointing out the inability of perceived need to increase helping can actually exacerbate the problem (Small, Loewenstein, & Slovic, 2007). As the authors of this work point out, emotion is often an essential component of decisions to help, and sometimes emotion can lead us to help in less than efficient ways (e.g., helping a single identifiable victim, rather than statistical victims in relatively greater need). Emotions can sway our beliefs when reasoned arguments cannot (DeSteno et al., 2004), and they guide our moral judgments even when we cannot explain why (Haidt, 2003). Considering the power of emotions to influence our notions of right and wrong, emotion seemed like an auspicious starting point for studying drivers of prosocial lying.

I will not repeat here the reasoning why compassion in particular is relevant to the study of prosocial lying (see "Benefits and Limitations of Compassion" section below). However, it became apparent through the development of this project and from the comments I received when presenting it that this research arguably forms a statement about compassion alongside

prosocial lying.

The illustration of compassion as a double-edged sword touches on Paul Bloom's (2017a, 2017b) critical examination of empathy as a tool for social good. According to Bloom, empathy should not be seen as virtuous given its tendency to be influenced by biases such as innumeracy (e.g., identifiable victim effect, Kogut & Ritov, 2005a; 2005b, Small & Loewenstein, 2003; 2005) and in-group favoritism (Hein et al., 2010). The in-group bias stemming from empathy, he argues, can be used to motivate war and other atrocities against out-groups. Bloom is very careful, as I am in this chapter, to distinguish between empathy and compassion (he defines compassion as caring about others without necessarily feeling their pain, and empathy as inferring what we think others are feeling). He claims that while we should de-emphasize the importance of empathy in promoting social good, compassion should be lauded. To support this idea, Bloom cites work providing evidence that while compassion predicts prosocial behavior, empathy does not (Jordan, Amir, & Bloom, 2016; Klimecki et al., 2013; 2014; Singer & Klimecki, 2014).

A critical issue with Bloom's analysis, however, is that compassion is subject to the same biases as empathy. In fact, recent work has shown that the identifiable victim effect occurs due to a lapse in compassion rather than empathy (Vastfjall, Slovic, Mayorga, & Peters, 2014). While other research on this effect has been less precise with terminology and assessment of mechanism, it is sympathy (i.e., compassion) that is proposed to bias people towards helping identifiable individuals when there are large numbers in greater need (e.g. Slovic, 2007; Small & Loewenstein, 2003). In-group bias, too, can be perpetuated by compassion (Cialdini et al., 1997). Thus, when I argue about the biased nature of compassion in this paper, I am really making the same argument as Bloom does for empathy. Bloom's failure to acknowledge that compassion

can bias decision making just as empathy does has recently been pointed out elsewhere (Vastfjall, Erlandsson, Slovic, & Tinghog, 2017). I view this paper through the lens of a biased compassion—an emotion that motivates people to help others, but does not necessarily provide an accurate map of how to best reach that goal.

Lying Because We Care:

Compassion Increases Prosocial Lying

When people are asked to report their most important moral value, the most frequent response is honesty (Graham, Meindl, Koleva, Iyer, & Johnson, 2015). Nevertheless, people report lying several times daily on average (DePaulo, Kashy, Kirkendol, Wyer, & Epstein, 1996). Many of these lies are told with the intention of benefiting others in some way, thus earning the classification "prosocial lie" (Levine & Schweitzer, 2014, 2015).

Despite the benevolent intentions behind prosocial lies, however, it is often the case that when given the opportunity to tell a prosocial lie, both lying and honesty can have different prosocial—and antisocial—consequences. For example, imagine a professor is asked by an undergraduate advisee to review his application essays for a prestigious doctoral program. After reading the essays, the professor thinks it unlikely that the student would be accepted into the program. Knowing that the student cares deeply about his academic identity and that he has put several months' effort into the materials, the professor believes the truth would be devastating to the student. At the same time, the professor understands that honest feedback will give the student an opportunity to revise the essays and significantly improve his chances at admission.

If the professor were to experience a rush of compassion for the student, how would it impact whether or not the professor gives the student honest feedback? One possibility is that compassion would lead the professor to consider the benefits of the honest feedback, and drive the professor to tell the student the hurtful but beneficial truth. That is, compassion could promote a focus on the student's career goals and help the professor see past the temporary emotional consequences of the feedback. Alternatively, compassion could instead focus the

professor on the negative emotional impact of the feedback, and lead the professor to tell a lie in the form of overly positive feedback.

In this paper, we explore, for the first time, the emotional basis of prosocial lying. Specifically, we examine how and why compassion impacts behavior when one has the opportunity to tell a prosocial lie. Determining how compassion influences prosocial lying is important for predicting the circumstances under which these lies might be told, as well as for developing an understanding of the counterintuitive and potentially detrimental effects of compassion on individuals, relationships, and organizations.

The Benefits and Limitations of Compassion

Compassion is sometimes confused with other related constructs in the empathy domain. Thus, we must first provide some conceptual work to make clear the construct we are studying. Under the superordinate heading of empathy lie three well-studied constructs (see Decety & Cowell, 2014; Levenson & Ruef, 1992; Preston & de Waal, 2002): (a) *Knowing what others feel* is a cognitive form of empathy that involves efforts to take the perspective of others (Zaki, 2014); success in this endeavor is called empathic accuracy (Ickes, 1993). (b) *Feeling what others feel* is an affective form of empathy that involves sharing the experiences of others (Wondra & Ellsworth, 2015), and is documented in rich literatures on emotional contagion (Barsade, 2002; Hatfield, Cacioppo, & Rapson, 1993; Neumann & Strack, 2000) and emotional mimicry (Dimberg, Thunberg, & Elmehed, 2000; Hess & Fischer, 2013). Finally, (c) *being emotionally motivated to alleviate others' distress or suffering* is an other-oriented emotion that involves an action tendency to help others (Goetz, Keltner, & Simon-Thomas, 2010); we label this construct "compassion" (see Haidt, 2003; Lazarus, 1991; Nussbaum, 1996), and study this

emotion in the present paper.[7] These three empathy-related constructs are psychologically (Davis, 1983; Konrath, O'Brien, & Hsing, 2011) and neurobiologically distinct (Decety, 2015; Immordino-Yang, McColl, Damasio, & Damasio, 2009; Shamay-Tsoory, Aharon-Peretz, & Perry, 2009), and predict different behavioral outcomes (Galinsky, Maddux, Gilin, & White, 2008; Jordan, Amir, & Bloom, 2016). However, there are some relations among the three constructs: for example, taking the perspective of a person in need promotes compassion (Coke, Batson, & McDavis, 1978), and empathic accuracy is facilitated by sharing others' feelings and physiological responses (Hess & Blairy, 2001; Levenson & Ruef, 1992).

Compassion is an emotion elicited by appraisals of need or undeserved suffering (Goetz et al., 2010; Haidt, 2003; Lazarus, 1991), and is often associated with increased prosocial behavior (Batson & Shaw, 1991; Eisenberg, 2002). Compassion is evoked by witnessing or learning about others' physical or emotional pain (Batson et al., 1997; Condon & DeSteno, 2011; Eisenberg et al., 1989; Stellar, Cohen, Oveis, & Keltner, 2014; Stellar, Feinberg, & Keltner, 2014; Stellar, Manzo, Kraus, & Keltner, 2012; Van Kleef et al., 2008) or victimization (Cameron & Payne, 2011; Valdesolo & DeSteno, 2011a), and by viewing depictions of suffering others such as homeless or malnourished people (Oveis et al., 2009; Oveis, Horberg, & Keltner, 2010). Philosophers and psychologists consider compassion to be the prototypical prosocial emotion, as it guides decisions about whom to help and how to help them (e.g., Cameron & Payne, 2012; Haidt, 2003; Nussbaum, 1996).

[7] Others have labeled this emotion as sympathy, empathy, or empathic concern (Batson, 1991; Batson & Shaw, 1991; Davis, 1983; Decety, 2015 Eisenberg, 1991; Eisenberg, 2002; Lazarus, 1991; Nussbaum, 1996; Wispé, 1986; see Haidt, 2003 for a discussion of construct terminology).

Because compassion involves appraisals of suffering in others, it is no surprise that this emotion increases prosocial behaviors aimed at alleviating suffering and harm. For example, participants induced to experience compassion become more willing to receive painful electric shocks in place of other people (see Batson & Shaw, 1991 for a review). In addition, participants who reported compassion while viewing footage of injured children offered to volunteer more time to help the family of those children (Eisenberg et al., 1989). Those experiencing compassion will help others even if they can escape the situation without doing so (Batson, Duncan, Ackerman, Buckley, & Birch, 1981). Compassion is also a motivator of generosity towards those who suffer (Saslow et al., 2013). Furthermore, nonverbal behaviors aimed to reduce suffering, such as soothing touch and skin-to-skin contact, have been observed cross-culturally (Hertenstein, Keltner, App, Bulleit, & Jaskolka, 2006).

Not only does compassion increase prosocial behaviors that involve *preventing* suffering and harm, but it also plays a role in behaviors that *promote* the welfare of others. When a person experiences compassion, their focus turns away from the goals and needs of the self and toward enhancing the welfare of others (Batson & Shaw, 1991; Eisenberg et al., 1989; Eisenberg & Miller, 1987; Horberg, Oveis, & Keltner, 2011; Oveis et al., 2010; Valdesolo & DeSteno, 2011b). As such, research suggests that compassion increases behaviors intended to help others, even at a cost to oneself. For example, compassion promotes forgiveness (Condon & DeSteno, 2011; Rudolph, Roesch, Greitemeyer, & Weiner, 2004), increases volunteerism (Omoto, Malsch, & Barraza 2009), and facilitates cooperation (Singer & Steinbeis, 2009).

Despite the multitude of work highlighting compassion's central role in prosocial behavior, however, researchers have recently begun documenting the limitations of compassion, as well as conditions under which this emotion can actually have perverse effects. An underlying

theme of this work is that compassion is associated with biases that can sometimes misguide our attention away from doing the "most good." This idea is well-illustrated by the story of Baby Jessica, who enraptured media attention and brought in hundreds of thousands of dollars in charitable donations after falling down a well, while elsewhere in the world, humanitarian crises such as the Kurdish genocide, which resulted in hundreds of thousands of lives being lost (Black, 1993), received little attention. Individuals experience more compassion towards identifiable victims than relatively greater numbers of victims described using statistics (Small & Loewenstein, 2003), and people downregulate their compassion when they encounter multiple victims in need because those needs appear overwhelming (Cameron & Payne, 2011). Compassion is also more easily and more often felt for those whose suffering is vivid (Loewenstein & Small, 2007), and in-group members, such as those who are closely related (Cialdini, Brown, Lewis, Luce, & Neuberg, 1997), or those who share our ethnicity or nationality (Stürmer, Snyder, Kropp, & Siem, 2006). It has been argued that the biased nature of compassion is a contributing factor to neglect of the world's greatest atrocities, the rectification of which requires overcoming of these biases so that people may recognize and act where help is needed most (Slovic, 2007).

Prosocial and Selfish Lying

Prosocial lying is ethically ambiguous. On one hand, lying violates the principle of honesty, a widely held moral value (Graham et al., 2015). Yet, these lies differ in their intentions from *selfish lies,* or those which are told to benefit oneself, potentially at the expense of others (Levine & Schweitzer, 2014). Selfish lies, such as those told for personal monetary gain, to protect one's status or position, or to attain social approval, are commonly viewed as reprehensible (Buller & Burgoon, 1994). In contrast, prosocial lies are colored by people's good

intentions, such as to prevent others from feeling hurt or embarrassed (DePaulo et al., 1996), or to benefit others financially (Erat & Gneezy, 2012).

It is important to note, however, that prosocial lies are benevolent in *intent*, but not necessarily in their ultimate consequences. That is, although those who tell prosocial lies have good intentions, these lies can have harmful effects on others. Providing overly positive feedback (such as in the professor-student example earlier) is one such context in which prosocial lies can ultimately backfire. Inflated feedback can harm performance (Ellis, Mendel, & Aloni-Zohar, 2009) and lead to avoidance of challenges (Brummelman, Thomaes, Orobio de Castro, Overbeek, & Bushman, 2014). Conversely, research has documented clear benefits to receiving accurate performance feedback. Accurate feedback can foster motivation to achieve goals and improve performance (Hyland, 1988; Ilgen, Fisher & Taylor, 1979; Locke & Latham, 1990). Research in organizational behavior has demonstrated the importance of accurate feedback for workplace productivity (Hillman, Schwandt, & Bartz, 1990), as well for clarifying expectations and reducing employee uncertainty (Ashford & Cummings, 1983). Thus, while prosocial lies are intended to benefit others, they may ultimately have detrimental effects on individuals and organizations.

Because of the adverse consequences that can result from prosocial lies, scholars across several domains of psychology (social, developmental, organizational behavior) and behavioral economics have sought to better understand these lies through research. One clear finding is that prosocial lying is ubiquitous. Prosocial lying is socialized early in life; parents lie to their children to promote positive emotions (Heyman, Luu, & Lee, 2009), and children in turn understand and tell prosocial lies themselves (Broomfield, Robinson, & Robinson, 2002; Talwar et al., 2007). Adults also tell prosocial lies regularly, especially in close relationships (DePaulo &

Kashy, 1998). Recent research has focused on responses to prosocial lying: Whereas selfish lies generally lead to distrust of the liar, prosocial lies that provide clear economic benefits to the target of the lie (hereafter "target") can increase trust and positive moral evaluations of the liar (Levine & Schweitzer, 2014, 2015). Yet, when the benefits of lying do not clearly outweigh those of honesty in the eyes of the target, prosocial lies can harm trust and moral judgments, and communicating benevolent intent may do little to mitigate these negative effects (Lupoli, Levine, & Greenberg, 2018). Other work has focused on predictors of prosocial lying: Research reveals that people are more likely to lie when others stand to gain (Gino, Ayal, & Ariely, 2013; Gino & Pierce, 2009; Wiltermuth, 2011), and prosocial lying is observed even when there is a cost to the self (Erat & Gneezy, 2012). Thus far, however, no work has examined what is likely a critical antecedent of prosocial lying: emotion, and in particular, the emotion of compassion.

Compassion and Prosocial Lying

Considering that compassion facilitates prosocial behavior, it seems likely that compassion would play *some* role in prosocial lying. What complicates matters, however, is that prosocial lying may not necessarily be the most beneficial action to take when considering targets' interests, because the alternative to prosocial lying might be helpful to them as well. When faced with the opportunity to tell a prosocial lie, two prosocial ethics are pitted against one another. Individuals must either lie in order to reduce harm or provide care to another, or tell the truth, which could also provide benefits for the target. Thus far, it is unclear how compassion influences behavior in moral dilemmas when different prosocial values are in conflict. In what direction might compassion influence prosocial lying, if any? Answering this question is critical to understanding compassion's influence on moral behavior, and this knowledge could inform policy initiatives aimed at increasing compassion in society and in organizations (e.g., Rynes,

Bartunek, Dutton, & Margolis, 2012).

On one hand, compassion could decrease prosocial lying (and thus produce increased honesty) for two reasons. First, when faced with the opportunity to tell a prosocial lie, those experiencing compassion might consider what is in the overall best interest of the target. As noted earlier, compassion has been shown to result in both harm-preventing behaviors, as well as behaviors that promote the wellbeing of others in ways unrelated to suffering. While no work has addressed how compassion influences behavior when harm prevention and non-harm-related welfare promotion are in conflict, one possibility is that compassion leads individuals to do whatever provides the greatest magnitude of benefits for others. Thus, if the benefits of a hurtful truth clearly outweigh the temporary pain inflicted by the truth, compassion could then lead an individual to be more honest. Recall the aforementioned example of the professor asked to evaluate the student's essays: Although hearing that that he is unlikely to be accepted would be painful, this would be a small price if honest criticism helps the student improve his application and ultimately gain admission. A compassionate individual might then be honest with the student about the flaws in his application.

Second, because lies have damaging effects on relationships, compassion may make individuals averse to telling lies in general. Deception can harm relationships by decreasing liking (Tyler, Feldman, & Reichert, 2006), intimacy (DePaulo et al., 1996), and trust (Schweitzer, Hershey, & Bradlow, 2006), and can also provoke revenge (Boles, Croson, & Murnighan, 2000). Additionally, in close relationships, such as friendships and romantic relationships, there are strong expectations of honesty (Stiff, Kim, & Ramesh, 1992). The discovery that one has been lied to can have negative emotional effects on the lie recipient, and damage or destroy the relationship (Haselton, Buss, Oubaid, & Angleitner, 2005; McCornack &

Levine, 1990). It is possible that a lifetime of exposure to the harmful consequences of lying in general could have spillover effects towards perceptions of prosocial lying. Thus, one experiencing compassion might opt to uphold the social contract of honesty, in part because of the detrimental effects that lying could have on one's relationships.

On the other hand, because compassion involves a heightened sensitivity to the suffering of others, this emotion could increase prosocial lying by focusing individuals on the harm inherent in a painful truth. That is, if lying is seen as a means to prevent or decrease suffering, then compassion might increase this type of lying. Consistent with this analysis is aforementioned work showing that compassion's effects on prosocial behavior are not necessarily calibrated toward promoting the *most* welfare-enhancing behavior, but instead toward promoting the welfare of others whose suffering is vivid (Loewenstein & Small, 2007). The circumstances under which lies are told lend well to compassion's biases: Lies are often told face-to-face, whereby the target is identifiable (e.g., Small & Loewenstein, 2003), and the pain that might result from the truth would be immediately apparent (i.e., vivid) to the potential deceiver. If the perceived harm that honesty might cause to the target is to be experienced in the here-and-now, compassion could act as a catalyst for prosocial lying in order to avoid this harm.

The Present Studies

In three studies, we provide the first tests of the influence of compassion on prosocial lying. We approach compassion at three levels (Han, Lerner, & Keltner, 2007; Rosenberg, 1998): as an experimentally-induced state experienced toward the potential target of a prosocial lie, or *integral compassion*; as an enduring emotional *trait*; and as an experimentally-induced state elicited by stimuli unrelated to the potential target of a prosocial lie, or *incidental compassion*. We also test whether a particular cognitive mechanism concerning the welfare of others—the

importance placed on preventing harm—might underlie the relationship between compassion and prosocial lying. Studies 1 and 2 examine prosocial lies that prevent emotional harm; Study 3 examines lies that promote the gains of others, while also investigating compassion's influence on selfish lies. All three studies measure real behavior. For all studies, we report all measures, conditions, and data exclusions.

Study 1:

Integral Compassion Increases Prosocial Lies That Prevent Emotional Harm

Study 1 tested whether experimentally-induced compassion (versus neutral feelings) would influence prosocial lying. Prosocial lying was operationalized as the inflation of feedback to the writer of a poorly written essay, as compared to participants' previous, private evaluations of that same essay. This behavioral paradigm simulates a regular occurrence in schools and workplaces in which individuals first evaluate an underperforming individual and then must decide whether to give accurate feedback.

Study 1 employed an integral manipulation of compassion; that is, the person who elicited compassion in the participants was also the potential target of the prosocial lie. This type of manipulation allowed us to examine compassion's relation to prosocial lying as it often occurs in the real world. We also tested a potential cognitive mechanism of compassion's influence on prosocial lying in this context—an enhanced importance placed on preventing harm to others, which is a primary appraisal of compassion (Goetz et al., 2010)—as well as potential alternative mechanisms.

Additionally, we included several measures to rule out alternative hypotheses that could potentially account for the effect of compassion on prosocial lying (if any). For instance, while some individuals respond to others' suffering with the other-oriented emotion compassion, which

predicts prosocial behavior, others experience personal distress, which is a self-focused response captured in measurements of one's own distress and anxiety, and does not predict prosocial behavior (Batson, 1991; Eisenberg et al., 1989; Eisenberg & Eggum, 2009; Eisenberg & Fabes, 1990). As such, we measured participants' emotional experience to determine whether the effect of compassion on prosocial lying (if any) was driven by compassion specifically, and to rule out the possibility that other affective responses—personal distress, other discrete emotions, positive affect, and negative affect—could explain the effects. Lastly, we measured social perceptions of the essay writer that could potentially account for the effect of compassion on prosocial lying.

Methods

Participants, design, and procedure. Participants were 434 undergraduates from a large U.S. public university. Participants were randomly assigned to the compassion or neutral condition in a two-cell between-subjects design. Twenty-four participants were excluded for failing an attention check, and nine participants were excluded for reporting suspicion that they were not actually paired with another individual. Five responses were excluded from individuals who had already participated in the study. This left a final sample of 396 participants (M_{age} = 21.3, 55.1% female), which fell just below our a priori target sample size of 400 (200 per cell).[8] We chose this sample size as a number that would give us high power to detect a small-to-medium effect size, given we did not have sufficient precedent to estimate a precise effect size.

[8] Twenty-eight of the respondents who were excluded were in the compassion condition, and 11 were in the compassion condition. The results of this study hold with the inclusion of all participants.

Participants completed the prosocial lying task (which included the compassion versus neutral manipulation), provided reports on their experienced emotions, and answered questions to assess potential mechanisms. Finally, we measured social perceptions of the writer to rule out potential confounding variables.

Prosocial lying task. We adapted a behavioral measure of prosocial lying (Jampol & Zayas, 2016) in which participants first provided private ratings of an essay written by another individual. They then read about a recent experience in this individual's life, which served as our manipulation of compassion or neutral feelings toward the essay writer. Next, they received a cover story explaining that they would have the opportunity to give the writer feedback, and that this feedback could help the writer improve the essay and thus improve his/her chance to earn a prize (see details below in section entitled, "Assessment of prosocial lying"). Finally, participants evaluated the essay a second time on the same dimensions, except this time with the knowledge that their evaluations would be shared with the essay writer. This procedure is graphically depicted in Figure 2.1.

As in Jampol and Zayas (2016), participants were first told that they would be paired with a student from another university who had written an essay about why he/she should be admitted to a graduate program. Participants were told that the purpose of the task was to let the researcher know (1) the quality of the student's writing, and (2) whether the writing sample should be provided to students who are applying to graduate school as an example of good "off the cuff" writing—that is, writing not prepared in advance. To bolster the believability of the cover story and to increase the salience of an identifiable target, participants were provided with the student's initials ("CG") and a short introductory message from this ostensible partner. Participants were also provided with a description of criteria they would use to evaluate specific

essay attributes (i.e., focus, logic, organization, support, mechanics), and were given an example of a high quality essay. Participants then read and rated the essay, which was pretested to be of relatively low quality ($N = 36$, sample drawn from same student population; $M = 44.56$, $SD = 20.69$; 0 = *worst*, 100 = *best*).

Private essay evaluations. Participants first provided their private ratings of the essay. Participants rated *quality* by indicating how the essay ranks "in general, compared to the best writing from someone in your peer-group/students at your university" (0 = *worst*, 100 = *best*). Participants' ratings of the focus, logic, organization, support, and mechanics of the essay—five attributes that are important in good essay writing, which were defined for participants—were averaged to form an *attributes* score ($\alpha = .74$; 1 = *worst*, 5 = *best*). Lastly, participants provided their *recommendation* for the essay ("How likely would you be to recommend this essay as a good example of off the cuff writing for students preparing for graduate admissions?"; 1 = *very unlikely*, 7 = *very likely*). Attributes and recommendation scores were converted to percentage of maximum possible scores (Cohen, Cohen, Aiken, & West, 1999); these scores and the quality score (which was already on a 0 to 100 scale) were averaged to form a measure of *overall private evaluations* ($\alpha = .76$). At no point were participants told that the writer would learn their identity or view their evaluation; thus, they were free to give any ratings they wished without social repercussions.

Manipulation of compassion versus neutral feelings toward the essay writer. After providing their initial private essay evaluations, participants received the manipulation of compassion or neutral feelings toward the writer. This manipulation was implemented in the form of a message ostensibly written by the essay writer about an event that recently occurred in his/her life. To reduce the potential for demand effects that could arise from identification of the

purpose of this message, we told participants that they would receive this message because "we want to give you the chance to know him/her [the writer] better," and that "he/she [the writer] was not given any specific instructions about what type of event he/she should write about."

Participants randomly assigned to the compassion condition then read a short paragraph adapted from Stellar, Feinberg, and Keltner (2014) that depicted the experience of a family member's death (with intentional spelling and punctuation errors to match the writing quality of the essay):

> *I dont know if this will be interesting to you but the only thing I can think of is two days ago my cousin passed away. It was really hard for me since we were so close. I spent a lot of time with her when I was younger we were best friends as kids. After I found out I just came home and sat in my room for a while by myself, my whole body was tired and I just felt so drained. I haven't talked to anyone about it really… I just couldn't believe it I, I wish I had gotten a chance to talk to her one last time. She was a really great person and she was a really big part of my life.*

Participants in the neutral condition read a paragraph about an ordinary grocery shopping experience.

Assessment of prosocial lying. After receiving the emotion manipulation, participants were asked to provide feedback to the writer about the quality of his/her essay. To (a) make the benefits of honesty salient, and (b) reduce demand effects that might arise from the perception that participants were expected to inflate their shared evaluations, we presented the following explanation to participants before they provided their feedback:

> *Your feedback is important. Each writer in this project must decide whether they would*

like to rewrite their essay before submitting it into a contest in which they can win a small prize that we will hold at the end of the semester. So, the information that you provide will help the writer improve his/her essay.

Participants again rated the *quality* and *attributes* of the essay and provided their *recommendation* for the essay on the same scales described above, but this time they received an on-screen reminder that their essay ratings would be shared with the essay writer. Attributes and recommendation ratings were converted to percentage of maximum possible scores, and these scores were averaged along with the quality rating to form a measure of *overall shared evaluations* ($\alpha = .79$).

***Experienced emotions*.** After providing their shared evaluations, participants were asked to think back to the message they read about the recent experience in the writer's life (the emotion manipulation), and to indicate the extent to which they experienced several emotions while reading this message (1 = *very slightly or not at all*, 5 = *extremely*). Twenty of the items assessed were taken from the Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988), and three additional items were used to assess compassion ("compassionate," "sympathetic," "moved"; Oveis et al., 2010). The order of the emotion items was randomized for each participant. We calculated composite scores for positive affect (10 items: interested, excited, strong, enthusiastic, proud, alert, inspired, determined, attentive, active; $\alpha = .85$), negative affect (10 items: distressed, upset, guilty, scared, hostile, irritable, ashamed, nervous, jittery, afraid; $\alpha = .77$), and compassion (3 items, $\alpha = .89$). In addition, we calculated a composite score for personal distress using a subset of the negative affect items (5 items: distressed, upset, scared, nervous, afraid; $\alpha = .74$), following past work that has measured personal distress with similar items (Eisenberg et al., 1989).

Mechanism: Harm prevention. A primary appraisal associated with compassion is a heightened focus on the suffering of others. Thus, we hypothesized that compassion's influence on prosocial lying would be mediated by an enhanced desire to prevent emotional harm. To assess this mechanism, participants responded to the following prompt: "When you were giving feedback to the student with whom you were paired during the second round of grading, how important was it for you to prevent any emotional harm or negative feelings that might have occurred as a result of your feedback?" (1 = *not at all important*, 7 = *extremely important*).

We also assessed alternative potential mechanisms by asking participants to indicate on the same scale how important it was to "give honest feedback," and how important it was to "give feedback that would help the student improve his/her writing." All mechanism questions were presented in randomized order.

Social perceptions. Next, we measured several perceptions of the writer. Participants were first asked, "How optimistic would you be about CG's [the writer's] success as a future graduate student?" (1 = *not at all*, 7 = *very*). They then received a series of questions on the same 1 to 7 scale in the following format: "How ___ is CG?" Participants rated the writer on the following dimensions: smart, dominant, warm, agreeable, competent, confident, open, likeable, trusting, trustworthy."

On the next survey page, we asked participants to indicate their beliefs about the gender of the student with whom they were paired (1 = *the student was very likely to be female*, 2 = *the student was probably female*, 3 = *the student could have been male or female*, 4 = *the student was probably male*, 5 = *the student was very likely to be male*). Lastly, participants responded to several exploratory measures, which are reported in the Supplemental Material and do not moderate the results.

Results

Manipulation check. The compassion induction was successful: Participants in the compassion condition reported feeling more compassion (M = 3.15, SD = 1.03) than did those in the neutral condition (M = 1.46, SD = 0.64), $t(394) = 19.76$, $p < .001$, $d = 1.99$.

Overall levels of prosocial lying. The prosocial lying task successfully generated prosocial lying. To test this, we subtracted overall private evaluations from overall shared evaluations. We also subtracted private from shard evaluations on each of the three evaluation criteria (quality, attributes, recommendation). The higher each difference score, the more participants inflated their ratings when giving feedback to the writer. For all measures, the mean difference score for each evaluation criterion across conditions was positive, indicating that participants provided more positive evaluations when the writer would view those evaluations, compared to their private evaluations ($M_{overall}$ = +3.67, $SD_{overall}$ = 8.94; $M_{quality}$ = +2.95, $SD_{quality}$ = 9.43; $M_{attributes}$ = +0.10, $SD_{attributes}$ = 0.41; $M_{recommendation}$ = +0.33, $SD_{recommendation}$ = 0.74). Furthermore, t-tests revealed that each of these difference scores significantly differed from zero (ps < .001), thus enabling us to reject the null hypothesis that no prosocial lying occurred.

Compassion increased levels of prosocial lying. In this study, we hypothesized that compassion would increase overall prosocial lying, which was operationalized as the size of the essay rating inflation going from overall private to overall shared evaluations. To test this, we ran a mixed model ANOVA. We entered condition (compassion/neutral) as a between-subjects variable, time (overall private/overall shared) as a within-subjects variable, and their interaction. With this analysis, the interaction term is the focal term: This tests whether the mean difference going from private to shared evaluations differs as a function of the manipulation.

Consistent with our hypothesis, this interaction was significant, $F(1,394) = 13.70$, $p < .001$, $\eta^2_p = .03$. The compassion condition produced increased overall prosocial lying (that is, a bigger difference going from private to shared evaluations) than the neutral condition ($M_{compassion} = +5.37$, $SD_{compassion} = 9.23$ vs. $M_{neutral} = +2.09$, $SD_{neutral} = 8.38$), $t(394) = 3.70$, $p < .001$, $d = .37$. There was also a significant main effect of time, $F(1,394) = 69.00$, $p < .001$, $\eta^2_p = .15$. Participants rated the essay higher when their evaluations were shared ($M = 29.68$, $SD = 16.36$) compared to than when they were private ($M = 26.01$, $SD = 14.78$), $t(395) = 8.18$, $p < .001$, $d = .41$. There was no main effect of condition ($p > .25$). These results are displayed in Figure 2.2.

We also examined prosocial lying on each of the three specific essay rating criteria (quality, attributes, recommendation) by running separate mixed model ANOVAs with each criterion entered as the dependent variable. Each of these models revealed significant interactions (quality: $F(1,394) = 15.21$, $p < .001$, $\eta^2_p = .04$; attributes: $F(1,394) = 8.19$, $p < .001$, $\eta^2_p = .02$; recommendation: $F(1,394) = 15.21$, $p < .001$, $\eta^2_p = .02$). Those in the compassion condition exhibited greater levels of prosocial lying (i.e., shared – private evaluations) in their ratings of quality ($M_{compassion} = +4.83$, $SD_{compassion} = 9.18$ vs. $M_{neutral} = +1.20$, $SD_{neutral} = 9.33$), $t(394) = 3.90$, $p < .001$, $d = .39$, attributes ($M_{compassion} = +0.16$, $SD_{compassion} = 0.37$ vs. $M_{neutral} = +0.05$, $SD_{neutral} = 0.29$), $t(394) = 2.86$, $p < .01$, $d = .29$, and recommendation ($M_{compassion} = +0.43$, $SD_{compassion} = 0.80$ vs. $M_{neutral} = +0.24$, $SD_{neutral} = 0.27$), $t(394) = 2.66$, $p < .01$, $d = .27$. In addition, these models revealed main effects of time (ps $< .001$), indicating that participants' shared ratings evaluations were significantly higher than their private evaluations (quality: $M_{private} = 33.88$, $SD_{private} = 20.36$ vs. $M_{shared} = 36.83$, $SD_{shared} = 20.95$; $t(395) = 6.24$, $p < .001$, $d = .31$; attributes: $M_{private} = 2.31$, $SD_{private} = 0.63$ vs. $M_{shared} = 2.41$, $SD_{shared} = 0.70$; $t(395) = 4.86$, $p < .001$, $d = .24$; recommendation: $M_{private} = 1.68$, $SD_{private} = 1.09$ vs. $M_{shared} = 2.02$, $SD_{shared} =$

1.19; $t(395) = 9.01$, $p < .001$, $d = .45$). There were no main effects of condition (ps > .25). Raw score means and standard deviations for private and shared evaluations across conditions are displayed in Table 2.1.

Importance placed on harm prevention partially mediated the effect of compassion on prosocial lying. After establishing that the compassion induction significantly increased prosocial lying, we assessed whether compassion also increased the importance placed on preventing emotional harm or negative feelings. Indeed, those in the compassion condition reported a significantly greater importance placed on preventing emotional harm than those in the neutral condition, $B = .39$, $p = .02$. The importance placed on preventing emotional harm also significantly predicted overall prosocial lying, $B = .65$, $p = .01$. We therefore examined the relationship between this potential mediator and overall prosocial lying. All mediation models implemented a difference score as the dependent variable, where overall private evaluations were subtracted from overall shared evaluations to obtain a measure of overall prosocial lying.[9]

Using the bootstrapping method, a mediation model with 20,000 bootstrap resamples confirmed that the importance placed on preventing emotional harm was a partial mediator of the relationship between compassion and overall prosocial lying, $B = .21$, 95% CI [.02, .59]. In

[9] F and p values for the Time x Manipulation interaction term in the mixed model ANOVA we reported are equivalent to F and p values for the independent variable in a one-way ANOVA where the manipulation (compassion/neutral) is the independent variable and the shared – private difference score is the dependent variable (Huck & McLean, 1975); both of these terms test whether the mean change going from private to shared evaluations differs as a function of the manipulation.

contrast, neither the importance placed on giving honest feedback nor the importance given to helping the student improve his/her writing was predicted by the compassion induction (*p*s > .25), thus ruling these items out as mediators of the relationship between compassion and prosocial lying.

Experienced compassion mediated the effect of the compassion manipulation on prosocial lying. In order to establish that the observed effects on prosocial lying were driven by the experience of compassion and not some other difference between the two experimental conditions, we first tested whether prosocial lying was predicted by experienced compassion as measured by the manipulation check. Overall prosocial lying was significantly predicted by experienced compassion, $B = 2.10$, $p < .001$. This effect held for both participants in the compassion condition, $B = 2.22$, $p < .001$, as well as those in the neutral condition, $B = 2.36$, $p < .01$. We also tested whether the data were consistent with a mediation model in which the experience of compassion mediates the influence of the compassion (versus neutral) condition on prosocial lying. The data were indeed consistent with such a model: A mediation model with 20,000 bootstrap resamples and bias-corrected confidence estimates revealed a significant indirect effect of the manipulation through experienced compassion on prosocial lying, $B = 3.81$, 95% CI [1.93, 5.96].

In addition, we tested multiple mediation models containing experienced compassion and other emotions as measured by items of the PANAS scale as mediators of the effect of the compassion manipulation on prosocial lying. A model containing experienced compassion, positive affect, negative affect, and personal distress as mediators revealed a significant indirect effect of experienced compassion, $B = 3.48$, 95% CI [1.09, 5.89], while confidence intervals around the indirect effects of positive affect, negative affect, and personal distress all contained

zero. These analyses serve as a test of the specificity of the effect, indicating that increases in prosocial lying stemmed from participants' experience of compassion, rather than other emotions.

Controlling for positive affect, negative affect, personal distress, specific emotions, and social perceptions did not account for the observed effects. The effect of the compassion manipulation on overall prosocial lying remained significant in a model controlling for positive affect, negative affect, and personal distress, $B = 2.14$, $p < .05$, and marginally significant in a model controlling for every specific emotion item assessed in the PANAS, $B = 2.00$, $p = .06$.

In addition, we looked for differences in social perceptions resulting from the compassion and neutral manipulation to determine if they could explain the effects on prosocial lying. Overall, those in the compassion condition ($M = 3.40$, $SD = 1.42$) reported being more optimistic about the writer's future as a graduate student than those in the neutral condition ($M = 2.93$, $SD = 1.34$), $t(394) = 3.40$, $p < .001$, $d = 0.34$. The writer in the compassion condition was also perceived as significantly more warm, agreeable, competent, open, likeable, trusting, trustworthy, and more likely to be female compared to the neutral condition (ps $< .05$). There were no significant differences between the two conditions in perceptions that the writer was smart, dominant, or confident (ps $> .20$). Importantly, the effect of the compassion manipulation on prosocial lying remained significant in a model controlling for each of the social perceptions significantly predicted by the compassion manipulation, $B = 3.73$, $p < .001$. Furthermore, a multiple mediation model with these perceptions entered as mediators revealed no significant indirect effects (all confidence intervals contained zero). We also ran a model controlling for positive affect, negative affect, personal distress, and the aforementioned social perceptions that

were influenced by compassion; the effect of compassion on prosocial lying remained significant in this model as well, $B = 2.79$, $p < .01$. [10]

Discussion

Study 1 provided the first demonstration that compassion increases prosocial lying. By examining peer feedback, the experimental design in this study simulated a common context in which prosocial is likely to occur. Moreover, we identified a mechanism: The effect of compassion on prosocial lying was partially mediated by the importance placed on preventing emotional harm that could occur as a result of their feedback. Other emotions and social perceptions of the target did not drive the effect.

Study 2:

Trait Compassion Predicts Increased Prosocial Lying To Prevent Emotional Harm

Study 2 tested whether individual differences in trait compassion predict prosocial lying using the same feedback paradigm implemented in Study 1. Trait emotions are enduring aspects of a person's personality that show stability over time and reflect elevated baseline levels of an emotion, increased tendencies to experience an emotion, and/or a decreased threshold for triggering the experience of an emotion (Rosenberg, 1998; Shiota, Keltner, & John, 2006). Investigating trait compassion thus offers another important glimpse into how prosocial lying effects are likely to emerge in the real world.

Methods

[10] Models that included covariates to rule out alternative hypotheses were linear mixed effects models with a random intercept for participant ID to control for repeated measures of private/shared ratings. Full regression tables are available in the Supplemental Material.

Participants, design, and procedure. Participants were 145 Amazon Mechanical Turk (Mturk) workers located in the United States. Four participants were excluded for failing an attention check, and two participants were excluded for reporting disbelief that they were paired with another individual. This left a final sample of 139 participants (M_{age} = 35.5, 60.5% female).[11] Before collecting data, we aimed to acquire as many participants as possible while staying within a budget.

No variables were manipulated in Study 2, thus eliminating the potential for demand characteristics that could arise from identification of the experimental manipulation. All participants completed the assessment of trait compassion, a filler task, the prosocial lying task, and the mechanism measures, as detailed below.

Trait compassion. Trait compassion was measured using two validated scales administered in counterbalanced order: the Empathic Concern subscale of the Interpersonal Reactivity Index (IRI-EC; Davis, 1983) and the compassion subscale of the Dispositional Positive Emotion Scales (DPES; Shiota et al., 2006). For the 7-item IRI-EC, participants indicated their agreement or disagreement (1 = *strongly disagree*, 5 = *strongly agree*) with items such as, "Other people's misfortunes usually do not disturb me a great deal," (reverse-scored) and "I often have tender, concerned feelings for people less fortunate than me." Internal reliability was high (α = .88). For the 5-item Compassion DPES, participants rated their agreement or disagreement (1 = *disagree strongly*, 7 = *agree strongly*) with items such as "Taking care of others gives me a warm feeling inside," and "I am a very compassionate person." Internal reliability was also high for this scale (α = .88). As expected, the two scales

[11] The results of this study hold with the inclusion of all participants.

were highly correlated ($r(137) = .86$), so we converted them to percentage of maximum possible scores and averaged them to form the composite measure of trait compassion ($\alpha = .92$).

Filler task and demographics. In order to disguise our hypotheses and preclude the desire for consistent responding with the trait compassion measures, it was important to temporally separate the compassion measures from the focal dependent variables. Thus, we provided participants with filler measures after assessing trait compassion. Here, participants answered demographic questions, then engaged in a task in which they formed neutral sentences from a series of scrambled words.

Prosocial lying task. We used the prosocial lying task from Study 1, with the cover story adapted for Mturk participants. Specifically, participants were told that we were interested in assessing Mturk workers' (those who participate in tasks on Mturk) perspectives on Mturk workers' writing. Participants were informed that they would be paired with another Mturk worker, and that this worker had been asked to write a short essay about the benefits of Mturk for both workers and requesters (those who post tasks on Mturk). As in Study 1, participants were informed that the purpose of the task was to let the researcher know the quality of the writing, and also to determine whether the essay should be included in an introductory manual for people potentially interested in using Mturk.

Similarly to Study 1, participants were shown the Mturk worker's initials and short introductory message. They then learned about the same criteria for evaluating specific essay attributes that were used in Study 1 (i.e., focus, logic, organization, support, mechanics). Next, participants provided private evaluations of the essay, which was rated in a pretest by Jampol and Zayas (2016) to be of low quality ($M = 22.20$, $SD = 19.20$ on a 0 [*worst*] to 100 [*best*] scale). The evaluation measures implemented here were also similar to those used in Study 1, with

minor changes. In Study 2, all measures were assessed on 0 to 100 scales. Participants rated the *quality* of the essay (0 = *worst*, 100 = *best*), the five essay *attributes* (0 = *worst*, 100 = *best*; α = .74), and the degree to which they would recommend the essay to be published in an introductory manual for online research (*recommendation*; 0 = *very unlikely*, 100 = *very likely*). Ratings on each criterion were averaged to form a measure of *overall private evaluations* (α = .89). The essay was provided on the screen while participants made their ratings.

After giving their initial, private evaluations, participants were asked to provide feedback to the writer about the quality of his/her essay. Before they gave their feedback, we presented participants with a similar explanation from Study 1 for why they would provide the feedback—that is, that their feedback was important because it could help the writer improve his/her essay before submitting it "into a future HIT [survey on Mturk] in which they can earn a bonus [extra money]." As in Study 1, we presented this information in order to make the benefits of honesty salient and to reduce potential demand effects.

Participants then evaluated the essay on the same three measures as before, with the addition of an on-screen reminder that these ratings would be shared with the essay writer. These evaluations were averaged to form a composite of *overall shared evaluations* (α = .89).

Mechanism: Harm prevention. Following the prosocial lying task, we asked participants the same question from Study 1 to assess the hypothesized mechanism—an enhanced focus on harm prevention—except that the writer was now referred to as a "worker" instead of a "student." Specifically, participants were asked, "When you were giving feedback to the worker with whom you were paired during the second round of grading, how important was it for you to prevent any emotional harm or negative feelings that might have occurred as a result of your feedback?" (1 = *not at all important*, 7 = *extremely important*). They were also asked the same

two questions from Study 1 to assess two alternative mechanisms: the importance placed on giving honest feedback, and on giving feedback that would help the worker improve his/her writing (1 = *not at all important*, 7 = *extremely important*). Following the mechanism questions, participants responded to additional exploratory measures, which are reported in the Supplemental Material and do not moderate results.

Results

Overall levels of prosocial lying. Once again, the prosocial lying task resulted in prosocial lying. Positive difference scores for overall prosocial lying as well as each evaluation criterion indicated that participants inflated their ratings when they would be shared with the writer, compared to their private evaluations ($M_{overall}$ = +3.51, $SD_{overall}$ = 7.55; $M_{quality}$ = +3.25, $SD_{quality}$ = 10.84; $M_{attributes}$ = +1.08, $SD_{attributes}$ = 7.87; $M_{recommendation}$ = +6.19, $SD_{recommendation}$ = 11.06). Additionally, t-tests revealed that difference scores for quality and recommendation measures significantly differed from zero (*p*s < .001), though difference scores for the attributes measure did not differ significantly from zero (p = .11).

Trait compassion predicts increased prosocial lying. To test our main hypothesis, we first examined correlations between trait compassion and overall prosocial lying, which was defined as the difference score of overall shared evaluations – overall private evaluations. Because the distributions of trait compassion scores were skewed (most participants rated themselves as relatively high in compassion (M = 75.28, SD = 15.72, Pearson's moment correlation of skewness = -.73), we conducted non-parametric Spearman rank-order correlations. Consistent with our predictions, trait compassion was significantly correlated with overall prosocial lying, $\rho(137) = .18$, p = .03. We then examined how prosocial lying correlated with the three evaluation criteria that comprised the composite measure. These analyses revealed a

significant positive correlation between compassion and prosocial lying about essay quality, $\rho(137) = .18$, $p = .03$, and recommendation, $\rho(137) = .21$, $p = .01$. The relationship between trait compassion and prosocial lying about the essay attributes was not significant ($p > .25$).

We also conducted additional analyses to determine how individuals who were both high and low in trait compassion rated the essay for both private and shared evaluations. We defined high trait compassion as one standard deviation above the mean or greater on our measure of compassion, and low trait compassion was defined as one standard deviation below the mean or less. Those who were high in trait compassion provided an overall private rating of 44.14 (SD = 24.79), and an overall shared rating of 50.13 (SD = 25.21). Those who were low in trait compassion had an overall private rating of 36.54 (SD = 18.64), and an overall shared rating of 40.75 (SD = 19.15). Means and standard deviations of private and shared ratings on each individual criterion for those high and low in compassion are provided in Table 2.2.

Importance placed on harm prevention partially mediated the relationship between trait compassion and prosocial lying. The relationship between compassion and our hypothesized mediator—the importance placed on preventing emotional harm or negative feelings—was significant, $\rho(137) = .27$, $p < .01$. The relationship between importance placed on harm prevention and overall prosocial lying was also significant, $\rho(137) = .23$, $p < .01$. As such, we tested whether the importance placed on preventing emotional harm mediated the relationship between trait compassion and prosocial lying. Consistent with Study 1, a mediation model with 20,000 bootstrap resamples indicated that the desire to prevent harm was a partial mediator of this relationship, $B = .02$, 95% CI [.01, .05] (See Figure 2.3).

Unlike in Study 1, however, compassion also predicted the importance placed on helping the worker improve his/her writing, $\rho(137) = .23$, $p < .01$, and the importance placed on giving

honest feedback, ρ(137) = .19, p = .02. Prosocial lying was significantly predicted by the desire to provide honest feedback, ρ(137) = -.30, $p < .001$, and marginally predicted by the desire help the worker improve, ρ(137) = -.15, p = .07. Therefore, we ran a multiple mediation model examining all three of these potential mediators simultaneously. There was again a significant indirect effect of the importance placed on harm prevention, B = .02, 95% CI [.002, .04]. However, confidence intervals for the indirect effects of the importance placed on helping the writer improve and on being honest both contained zero, thus ruling these out as mediators of the relationship between trait compassion and prosocial lying.

Discussion

In Study 2, trait compassion predicted increased prosocial lying. While this study implemented a correlational design, the results are consistent with those of Study 1, thus offering more evidence for the positive relationship between compassion and prosocial lying. Further supporting this evidence is the identification of the same underlying mechanism in Studies 1 and 2. In both of these studies, the desire to prevent emotional harm partially mediated the relationship between trait compassion and prosocial lying, rather than alternative mechanisms.

Study 3:

Compassion Increases Prosocial Lies That Promote the Gains of Others But Not the Self

Whereas Studies 1 and 2 examined how compassion influences and relates to lies that *prevent* harm to others, Study 3 instead examined lies that *promote* positive outcomes for others. Specifically, Study 3 investigated whether experimentally-induced compassion would increase lies that procure financial gains of others—in this case, a charity. By examining prosocial lying in a different context, Study 3 helps to assess the external validity of the effects seen in Studies 1 and 2. Moreover, in this study, we examined a third form of compassion by testing the effect of

incidental state compassion on prosocial lying. That is, we manipulated compassion that was unrelated to the subsequent target of a prosocial lie. Testing the effects of incidental compassion on prosocial lying offers another key glimpse into how prosocial lying might unfold in the real world, as emotions can have spillover effects on decision-making in a variety of domains (e.g., Han et al., 2007). Lastly, we tested discriminant validity by investigating both prosocial and selfish lies, predicting moderation such that compassion would increase prosocial lies, but either decrease or have no effect on selfish lies.

Methods

Participants, design, and procedure. Participants were 455 undergraduates from a large U.S. public university. Participants were randomly assigned to one of four conditions in a 2 (Emotion: compassion/neutral) x 2 (Lie Type: prosocial/selfish) between-subjects design. Ten participants were excluded due to a computer malfunction, three were excluded for being familiar with the lying task, six were excluded for guessing the hypothesis of the study, and four were excluded for displaying consistent responding that demonstrated a lack of understanding or concern for the task (by giving the payoff-minimizing response for the first 100 trials of the task). This left a final sample of 432 (M_{age} = 21.3, 49.2% female).[12] Before collecting data, we had a target sample size of at least 400 (100 per cell), and planned to collect as many responses as possible within the lab time we were allotted to run the study. All participants received course

[12] Of those participants who were excluded, nine were in the compassion/prosocial condition, nine were in the compassion/selfish condition, three were in the neutral/prosocial condition, and one was in the neutral/selfish condition. The results of this study hold with the inclusion of all participants.

credit in exchange for participation; additional incentive payments were made to a random selection of 10% of participants according their responses in the lying task (it was possible to gain up to $10 in incentive payments for the self or for charity).

To obscure the study's purpose, participants were first told that they would be participating in a study about "how personality and visual stimuli influence memory." To bolster the cover story about the memory task, participants were told, Next, participants filled out the Big Five Personality Inventory (BFI; John, Donahue, & Kentle, 1991), which assessed control variables. Then, participants received the compassion or neutral emotion induction, completed the lying task (where lies benefited the self or others), and finally reported on their experienced emotions.

Big Five Personality Inventory (control variables). Participants completed the 44-item BFI on 1 (*strongly disagree*) to 5 (*strongly agree*) scales. We measured agreeableness as a control variable because of its potential relationship with decisions to lie prosocially, and because agreeableness, along with extraversion, tends to covary with positive emotionality (John & Srivastava, 1999). Neuroticism was measured as a control variable because of its empirical links with negative emotionality. We additionally included conscientiousness and openness to experience as control variables because they make up the other two major dimensions of personality.

Emotion manipulation – compassion vs. neutral. Next, participants received the emotion manipulation. Those in the compassion condition viewed a validated 15-slide compassion induction (photographs depicted helplessness and vulnerability; Oveis et al., 2010) followed immediately by a validated 46-second film induction of compassion (about child malnutrition and starvation; Côté et al., 2011). Importantly, the slides and video selected were

not connected to the target organization of the prosocial lying task, nor was it plausible based on photo/video content or procedure that participants would later believe that they were benefiting the individuals depicted in the compassion induction.

Participants in the neutral condition viewed 15 neutral slides from the International Affective Picture System (IAPS; Lang, Bradley, & Cuthbert, 1999) immediately followed by a 46-second clip from the film *All the President's Men* depicting two men talking in a courtroom—a clip that past research has shown to elicit a neutral state (Hewig et al., 2005). All stimuli used in the manipulation can be found in the Supplemental Material.

Lying task – prosocial lies vs. selfish lies. Immediately after the emotion induction, participants engaged in a lying task adapted from Gino, Norton, and Ariely (2010). For this task, participants viewed a series of arrays of dots dispersed within a square. Each square had a diagonal line cutting it in half, such that some dots were displayed to the right of the diagonal, and some dots to the left of the diagonal. After a 1-second exposure to each trial, participants were asked report whether there were more dots to the left or the right of the diagonal by pressing one of two keys.

Participants in the selfish lie condition were told that they would be paid 0.5 cents each time they reported that there were more dots on the left, and 5 cents for each time they reported that there were more dots on the right "because most people can easily identify the number of dots on the left side." That is, they were incentivized to say that there were more dots on the right regardless of whether or not this was true.

In the prosocial lie condition, participants received the same information, but were told that the money earned based on their responses would be donated to a real charity—the Against Malaria Foundation. Participants in this condition were also given a short paragraph about the

nature of the charity, which provides insecticide-treated mosquito nets for the prevention of malaria (see Supplemental Material for full description provided to participants). All money earned by participants in the prosocial lie condition was actually donated to the Against Malaria Foundation.

Following Gino et al. (2010), all participants first performed 15 practice trials. After the practice phase, there were 200 trials divided into two blocks with 100 trials each. Each of the two blocks contained 34 trials in which there were clearly more dots on the left (a right-to-left ratio of less than 2/3), 50 trials in which it was ambiguous whether there were more dots on the left or the right (a right-to-left ratio greater than or equal to 2/3 and less than or equal to 3/2), and 16 trials in which there were clearly more dots on the right (the ratio of the number of dots on the right to the number of dots on the left was greater than 3/2). As in Gino et al. (2010), *clearly dishonest* responses were defined as "more on the right" responses—the response that yielded the higher payoff—when there were clearly more dots on the left. *Ambiguously dishonest* responses were defined as "more on the right" responses when it was ambiguous whether there were more dots on the right or left. *Honest* responses were defined as "more on the right" responses when they were clearly more dots on the right.

Experienced emotions. Immediately following the lying task, participants completed the same measures of experienced emotions as in Study 1 for our manipulation check. Here, participants were asked to indicate the extent to which they experienced each emotion after viewing the slides and video. We once again calculated scores for positive affect (10 items, α = .89), negative affect (10 items, α = .90), personal distress (5 items; α = .85) and compassion (3 items, α = .90). All items were displayed in a randomized order. Due to a programming error, only 269 of the 432 participants were asked about their experienced emotions.

Results

Manipulation check. We ran a 2 (Emotion: compassion/neutral) x 2 (Lie Type: prosocial/selfish) ANOVA on experienced compassion as our manipulation check. As expected, there was no main effect of lie type ($p > .25$), but there was a significant main effect of emotion condition, $F(1,265) = 267.12$, $p < .001$, $\eta^2_p = .50$. The previously validated emotion induction successfully induced compassion: Participants in the compassion condition ($M = 3.38$, $SD = 0.98$) reported more experienced compassion than those in the neutral condition ($M = 1.62$, $SD = 0.82$), $t(267) = 16.06$, $p < .001$, $d = 1.96$. This analysis also revealed an unpredicted significant interaction, $F(1,265) = 11.84$, $p < .001$, $\eta^2_p = .04$. The compassion condition resulted in a greater increase in experienced compassion for those in the prosocial lie condition ($M = 3.60$, $SD = 0.96$ vs. $M = 1.47$, $SD = 0.70$, $t(132) = 14.85$, $d = 2.57$) than those in the selfish lie condition ($M = 3.15$, $SD = 0.96$ vs. $M = 1.76$, $SD = 0.90$, $t(133) = 8.66$, $d = 1.50$).

Prosocial and selfish lying. Overall, this procedure successfully produced prosocial and selfish lying. Those in the prosocial lie conditions exhibited on average 41.15 clearly dishonest responses ($SD = 14.66$) out of a potential 68 trials (60.51%), and 63.72 ambiguously dishonest responses ($SD = 18.50$) out of a potential 100 trials (63.72%). Those in the selfish lie conditions demonstrated on average 38.08 clearly dishonest responses ($SD = 13.31$) out of 68 trials (56.0%), and 60.13 ambiguously dishonest responses ($SD = 17.02$) out of 100 trials (60.13%).

For each dependent variable (clearly dishonest responses, ambiguously dishonest responses, honest responses), we conducted a 2 (Emotion: compassion/neutral) x 2 (Lie Type:

prosocial/selfish) ANOVA.[13] For ease of comprehension, for each dependent variable we used the percentage of dishonest responses, rather than the absolute number of dishonest responses.

For clearly dishonest responses, as predicted, there was a significant Emotion x Lie Type interaction, $F(1,428) = 6.51$, $p = .01$, $\eta^2_p = .01$ (see Figure 2.4, Panel A). Participants in the compassion condition ($M = 63.61$, $SD = 23.60$) exhibited more clearly dishonest responses for the benefit of the charity (i.e., prosocial lying) than those in the neutral condition ($M = 57.66$, $SD = 19.16$), $t(212) = 2.03$, $p = .04$, $d = .28$. There was no statistically significant difference in clearly dishonest responses for participants' own monetary gain (i.e., selfish lying) between those in the compassion condition ($M = 53.79$, $SD = 19.18$) and those in the neutral condition ($M = 57.91$, $SD = 19.78$), $p = .12$. In addition, there was a main effect lie type, $F(1,428) = 5.28$, $p = .01$, $\eta^2_p = .01$. Those in the prosocial lie conditions ($M = 60.52$, $SD = 21.56$) demonstrated more clearly dishonest responses than those in the selfish lie conditions ($M = 56.00$, $SD = 19.57$). There was no main effect of emotion ($p > .25$).

For ambiguously dishonest responses, similar results were obtained (see Figure 2.4, Panel B). As predicted, there was a significant Emotion x Lie Type interaction, $F(1,428) = 5.96$, $p = .02$, $\eta^2_p = .01$. Those in the compassion condition ($M = 66.78$, $SD = 20.29$) exhibited more prosocial lying than those in the neutral condition ($M = 60.89$, $SD = 16.26$), $t(212) = 2.35$, $p = .02$, $d = .32$. There was no statistically significant difference in selfish lying between those in the compassion condition ($M = 58.83$, $SD = 16.39$) and those in the neutral condition ($M = 61.26$, $SD = 17.54$), $p > .25$. There was also a main effect of lie type, $F(1,428) = 4.45$, $p = .04$, $\eta^2_p = .01$,

[13] Repeated measures analyses with block (first vs. second) included as a factor are included in the Supplemental Material, though inclusion of block as a factor does not alter the results.

such that participants engaged in more lying in the prosocial lie conditions (M = 63.72, SD =18.50) than in the selfish lie conditions (M = 60.14, SD = 17.02). There was no significant effect of emotion ($p > .25$).

For honest responses, as predicted, there was no significant Emotion x Lie Type interaction ($p > .25$; see Figure 2.4, Panel C). There was also no main effect of lie type ($p = .11$) nor emotion ($p > .25$).

Experienced compassion predicted prosocial lying. As an additional test of the specificity of the observed effects, we examined whether prosocial lying was predicted by experienced compassion, as measured by our manipulation check. Experienced compassion marginally predicted clearly dishonest responses, $B = 2.48$, $p = .07$, and significantly predicted ambiguously dishonest responses, $B = 2.44$, $p = .04$. However, experienced compassion did not mediate the effect of compassion on prosocial lying.

Controlling for positive affect, negative affect, personal distress, specific emotions, and personality traits did not account for the observed effects. To ensure that these effects were specific to compassion and were not due to other emotions or personality traits, we examined the effect of the compassion manipulation on prosocial lying with the inclusion of covariates to control for these other emotions and personality traits. The effect of compassion on prosocial lying (for both clearly dishonest and ambiguously dishonest responses) held in models controlling for positive affect, negative affect, and personal distress (clearly dishonest responses: $B = 9.14$, $p < .05$; ambiguously dishonest responses: $B = 10.23$, $p < .01$), as well as in models controlling for all individual items of the PANAS (clearly dishonest responses: $B = 16.09$, $p < .01$; ambiguously dishonest responses: $B = 16.22$, $p < .01$).

In addition, the effect of compassion on prosocial lying held in models simultaneously controlling for extraversion, agreeableness, neuroticism, conscientiousness, and openness (clearly dishonest responses: $B = 7.90$, $p < .05$; ambiguously dishonest responses: $B = 8.08$, $p < .05$). Lastly, we ran models examining the effect of compassion on prosocial lying controlling for personality traits, as well as positive affect, negative affect, and personal distress. The effect of compassion on prosocial lying also held in these models (clearly dishonest responses: $B = 8.84$, $p = .06$; ambiguously dishonest responses: $B = 10.05$, $p = .01$) Thus, enduring personality traits and other emotions did not account for the observed effects.[14]

Discussion

Consistent with Studies 1 and 2, Study 3 found that incidental compassion increased prosocial lying. Critically, the compassion-eliciting stimuli were unrelated to the charity that benefited from participants' dishonest behavior, and the compassion induction still increased prosocial lying.

These results expand the findings of Studies 1 and 2 in several ways. First, Study 3 employed a different operationalization of compassion, and also examined a different type of compassion. Using a large sample, we found that prosocial lying is not only associated with integral (Study 1) and trait (Study 2) compassion, but is also increased by incidental compassion (Study 3). These results offer further evidence for the causal influence of compassion on prosocial lying. Second, the use of another operationalization of prosocial lying in Study 3 bolsters support for the external validity of the effect. In addition to being associated with prosocial lying that prevents emotional harm in the context of providing performance feedback, compassion also increased prosocial lies that promoted financial benefits for a humanitarian aid

[14] Full regression tables for these models are available in the Supplemental Material.

charity. This phenomenon could present itself in the real world in the form of a charity employee lying on tax returns to reserve more funds for humanitarian work. Third, by examining two types of lies—selfish and prosocial lies—we demonstrated that the beneficiary of the lie is an important moderator of the relationship between compassion and deception. Compassion increased prosocial lying, but not selfish lying. Furthermore, we again ruled out important alternative explanations: Other emotions did not explain these effects, nor did personality traits linked to positive affect (extraversion and agreeableness), negative affect (neuroticism), or prosocial behavior (agreeableness).

General Discussion

The present studies provide the first investigation of the emotional underpinnings of prosocial lying. Across studies, we examined compassion at three different levels, demonstrating that both integrally (Study 1) and incidentally (Study 3) induced state compassion causally increase prosocial lying, and that individual differences in trait compassion (Study 2) are positively associated with prosocial lying. Not only did we implement multiple operationalizations of compassion, but we also studied two different types prosocial lies: those that prevent emotional harm, and those that promote the welfare of others. All studies investigated actual lying behavior, rather than attitudes toward lying or hypotheticals. Furthermore, we ruled out alternative explanations across studies that could potentially account for our results—that is, we found that the observed increases in prosocial lying were due to compassion specifically, and not due to other discrete emotions, personal distress, generalized positive or negative affect, personality traits, or social perceptions of the target. Together, this research demonstrates how compassion increases prosocial lying.

In addition to uncovering the relationship between compassion and prosocial lying, we also identified a mechanism behind this effect in Studies 1 and 2. In the context of providing feedback, the effect of compassion on prosocial lying was partially mediated by the importance placed on preventing emotional harm. Compassion has been shown to increase prosocial behaviors associated with both harm prevention (e.g., Batson et al., 1981) as well as non-harm-related welfare promotion (e.g., Condon & DeSteno, 2011). However, this mechanism suggests that compassion may make individuals particularly attuned to preventing the suffering of others, even when additional routes to helping others are available (e.g., providing honest feedback).

Moreover, in Study 3, we showed that compassion increased lies that helped a charity, but had no effect on lies that financially benefited participants themselves. This suggests that compassion does not exert global effects on deception, but rather that the beneficiary of the lie is an important moderator of the relationship between compassion and dishonesty. Although the present investigation is focused on how compassion influences prosocial lies, it is worth noting that, to our knowledge, these are the first data to investigate whether compassion influences selfish lies. Thus, while compassion may promote prosocial behavior, this emotion may not have any appreciable (negative) effect on antisocial behavior.

This work contributes to the nascent literature on prosocial lying in several ways. First, no research has examined emotion as a causal driver of prosocial lying. Previous research on prosocial lying has focused on identifying contexts in which these lies are told (e.g., DePaulo et al., 1996), responses to those who tell prosocial lies (e.g., Levine & Schweitzer, 2014), or qualitative assessments of reasons for lying (e.g., DePaulo & Kashy, 1998). Our research extends theory on prosocial lying by providing the first demonstration that compassion is related to and causally influences prosocial lying. In addition, this research provides insight into an important

real world context in which prosocial lies are told. Past work has often operationalized prosocial lying using economic games (e.g., Erat & Gneezy, 2012; Levine & Schweitzer, 2014, 2015), which afford experimental control but do not closely resemble real world situations in which lies are told. Given the usefulness of these games for cleanly differentiating prosocial lies from other types of lies (e.g., selfish lies), we borrowed from this approach for our lying task in Study 3. However, by examining prosocial lying in the form of overly inflated person-to-person feedback in Studies 1 and 2, we shed light on how compassion influences behavior in a common situation that affords the opportunity for prosocial lying.

This work also informs scholarly understanding of compassion and how it shapes ethical behavior. While compassion's positive influence on prosocial behavior has been widely documented, little work has examined how compassion affects moral decision making, and no work has examined how compassion influences behavior when different ethical principles are pitted against one another. According to Moral Foundations Theory (Graham et al., 2011; Haidt & Graham, 2007; see also Shweder, Much, Mahapatra, & Park, 1997), people across cultures conceive of actions and beliefs in several different domains as morally relevant. Lying may be regarded as a violation of the principle of honesty (Graham et al., 2015) and the decision to tell a prosocial lie presents a conflict between the principle of honesty and the principle of harm and care—the obligation to aid the welfare of others. Our work suggests that compassion might cause people to consider harm and care more heavily in ethically ambiguous situations. More research would help to illuminate how compassion influences the weighting of harm and care relative to other moral values across a broader spectrum of moral dilemmas.

In addition, this research contributes to a growing body of work that highlights how, despite the prosocial benefits it often affords, compassion can sometimes lead individuals to act

contrary to what is truly in others' best interests (e.g., Cameron & Payne, 2011; Slovic, 2007). Similarly to how compassion draws attention and resources to identifiable victims rather than to comparably greater atrocities (Small, Loewenstein, & Slovic, 2007), our results suggest that compassion may bias individuals toward alleviating immediate emotional harm rather than attending to others' longer-term goals (e.g., performance improvement resulting from critical feedback). This notion is consistent with work suggesting that affect and emotion play an important role in intertemporal choice (DeSteno, 2009; Hirsh, Guindon, Morisano, & Peterson, 2010; Loewenstein, 1996), and in (mis)predicting the preferences and emotions of others (Van Boven & Loewenstein, 2003). However, it may also be that when honesty is perceived to result in future benefits for a target that far outweigh the benefits of lying, compassion could lead individuals to be more honest. While recent work has begun to address how positive emotions such as gratitude influence temporal discounting (DeSteno, Li, Dickens, & Lerner, 2014; Dickens & DeSteno, 2016), further research is necessary to understand how compassion influences valuations of others' short-term and long-term goals.

Another area for future research lies in how the relationship between the lie teller and the target of the lie moderates the effect of compassion on prosocial lying. In the present studies, participants were given the opportunity to lie only to strangers. As such, it is critical to determine whether these effects generalize to closer relationships. The relationship between compassion and prosocial lying may differ depending on the in-group/out-group membership of the lie target, or the lie teller's perceived closeness to the target. People feel more compassion towards those to whom they are closely related (Cialdini et al., 1997), and people also tell more prosocial lies to close others than selfish lies (DePaulo & Kashy, 1998). Thus, it is possible that an interaction exists between compassion and the closeness of the lie target on prosocial lying, such that

compassion would exert an even stronger influence on prosocial lies told between friends, coworkers, or relationship partners.

One limitation of our studies is that we did not assess the extent to which participants considered their own behavior as dishonest. While it would be interesting to know whether individuals were consciously aware that they were lying, we would argue that conscious awareness is not a necessary condition for dishonesty. Individuals often lack conscious insight into their mental processes (Nisbett & Wilson, 1977), and self-deception is common (Mazar, Amir, & Ariely, 2008; Tenbrunsel & Messick, 2004). Furthermore, it is possible that even if participants did consider their behavior dishonest, that they would not admit this upon being asked due to social desirability concerns. We encourage future research to determine if people's conscious awareness of their dishonesty is a moderating factor in the relationship between compassion and prosocial lying.

It is also important to note that the mechanism uncovered behind the effects seen in Studies 1 and 2 does not apply to Study 3; that is, when lying for the financial gain of a charity, there is no emotional harm to be prevented. However, we believe a similar mechanism might underlie the results in Study 3, whereby importance is still placed on reducing harm, albeit not emotional harm. In the context of Study 3, dishonest responding could result in more money being donated to the Against Malaria Foundation for the purchase of mosquito nets to prevent the spread of malaria. Supporting this cause financially could thereby prevent harm and human suffering. Although we did not measure participants' views about the extent to which their actions in the task could reduce suffering, we speculate this belief could mediate the effect of compassion on prosocial lying for others' gains—a hypothesis worthy of further investigation.

According to Ralph Waldo Emerson (1888), "the purpose of life…is to be honorable, to be compassionate, to have it make some difference that you have lived and lived well." Unfortunately, Emerson did not offer guidelines for how one should behave when helping others requires an act that some may view as dishonorable, such as lying. The present research suggests that compassion may provide that moral compass by leading individuals to tell lies that are intended to benefit others. Indeed, many people likely lie not in spite of their concern for others, but rather because they care.

Chapter 2, in full, is a reprint of material as it appears in Journal of Experimental Psychology: General, which was co-authored by Lily Jampol and Christopher Oveis in 2017. The dissertation author was the primary investigator and author of this paper.

Figures

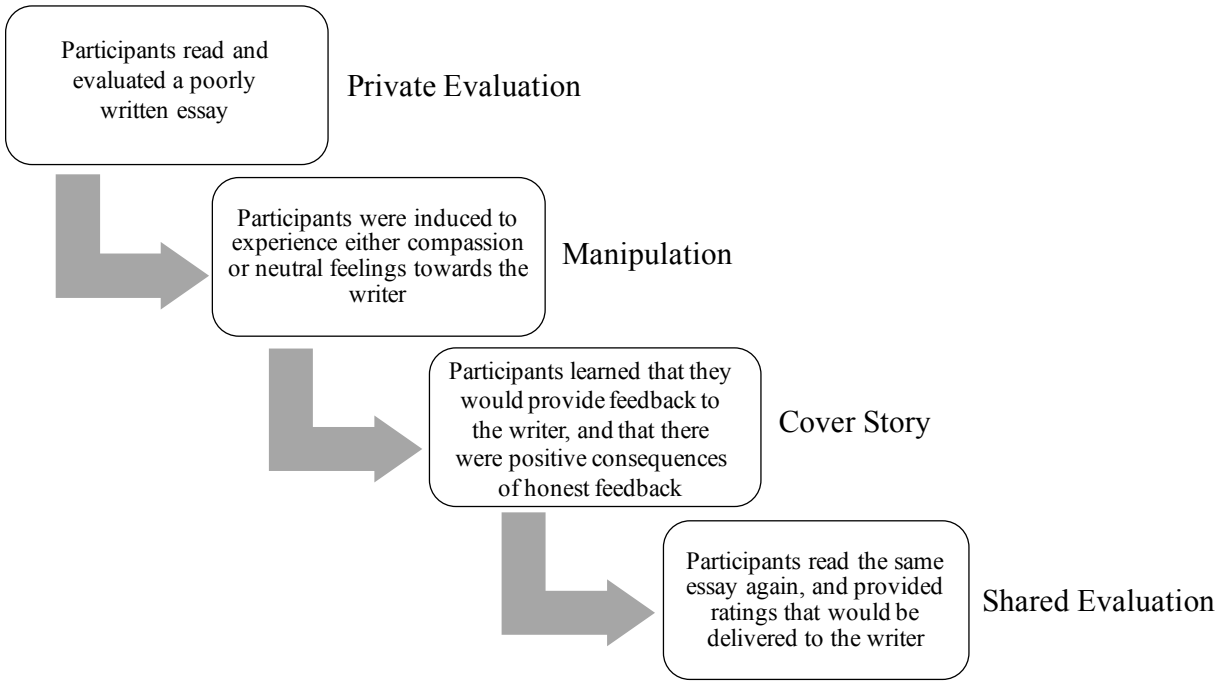


Figure 2.1: Overview of prosocial lying task in Study 1.

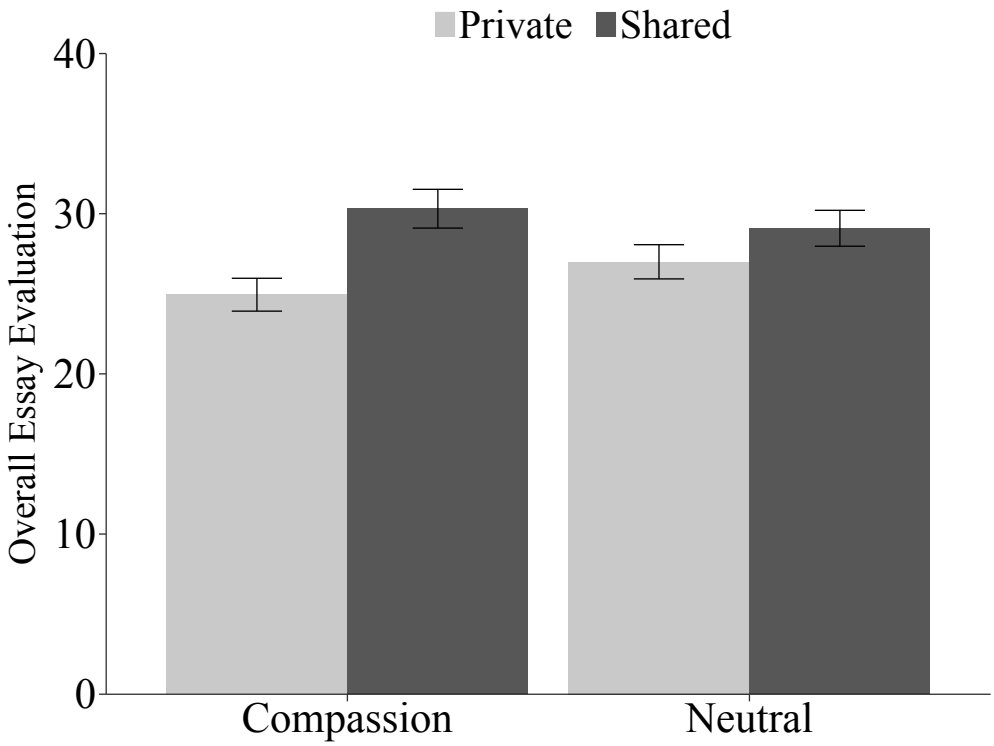
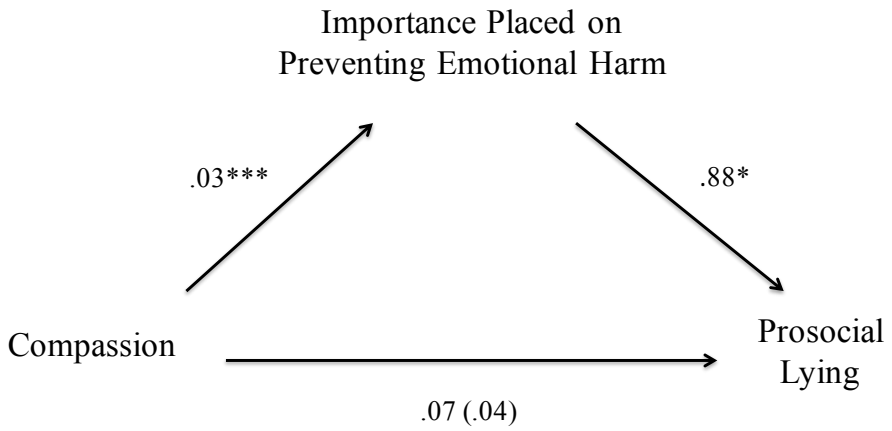


Figure 2.2: The effect of integral compassion on overall essay evaluations in Study 1. Essay evaluations are on a 0 to 100 scale. Error bars signify standard errors.



Indirect Effect: $B = .02, 95\%\ \text{CI}\ [.01, .05]$

Figure 2.3: The relationship between compassion and prosocial lying as mediated by the importance placed on preventing emotional harm. Trait compassion is on a 0 to 100 scale. Coefficient in parentheses represents the relationship between compassion and prosocial lying controlling for importance placed on preventing emotional harm. * $p < .05$, *** $p < .001$.

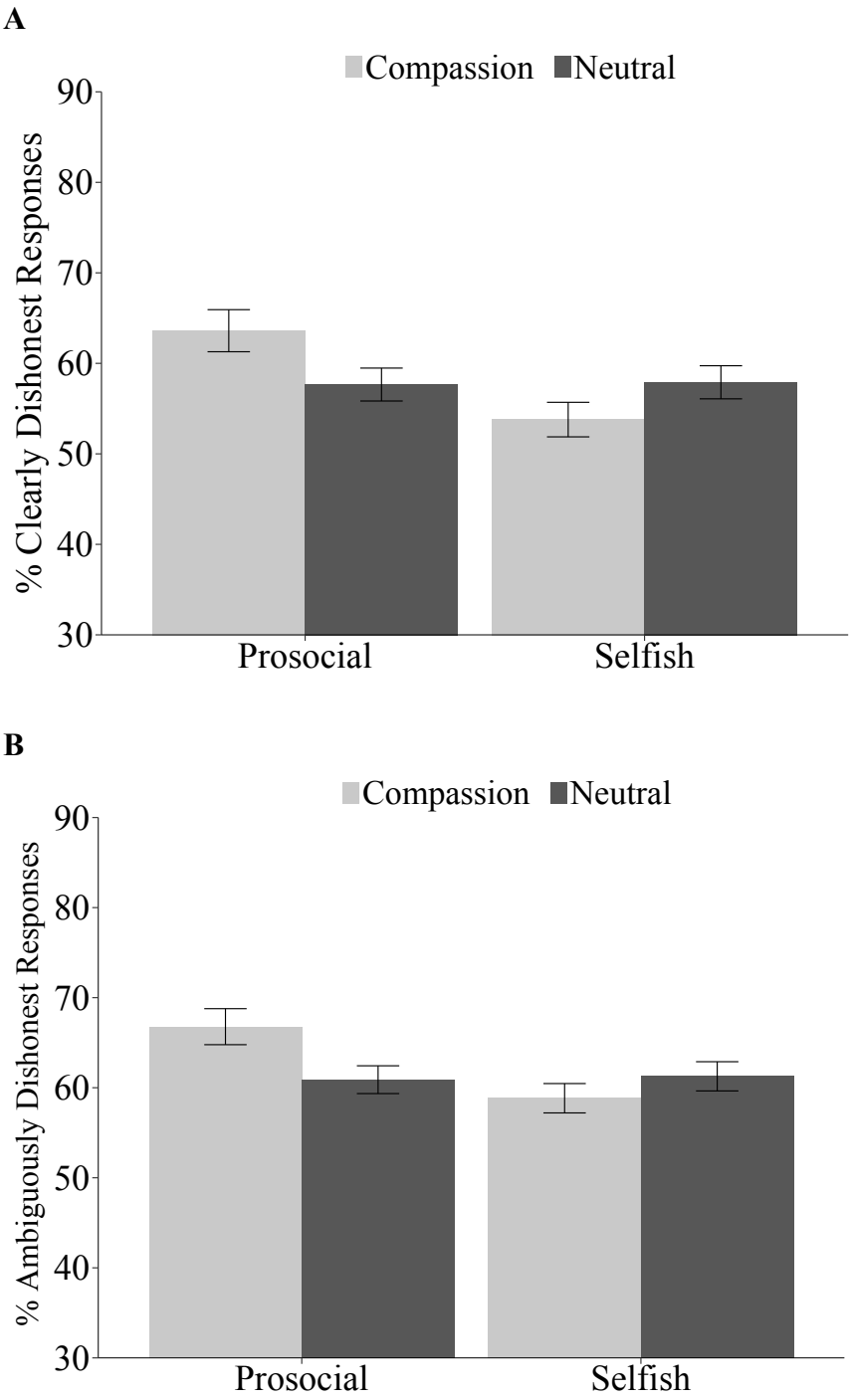


Figure 2.4: The effect of incidental compassion on clearly dishonest responses (Panel A), ambiguously dishonest responses (Panel B), and honest responses (Panel C) for prosocial and selfish causes in Study 3. Error bars signify standard errors.

C

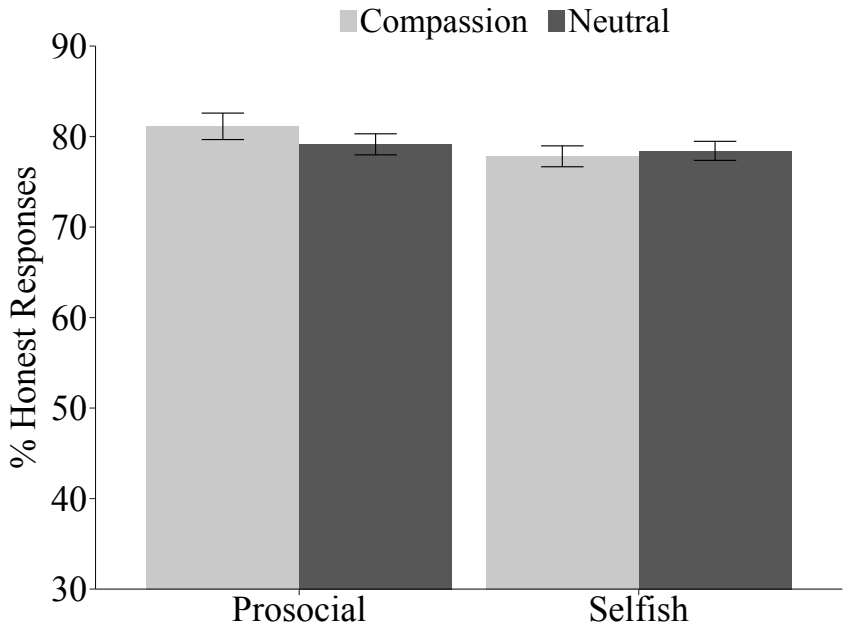


Figure 2.4, Continued: The effect of incidental compassion on clearly dishonest responses (Panel A), ambiguously dishonest responses (Panel B), and honest responses (Panel C) for prosocial and selfish causes in Study 3. Error bars signify standard errors.

Tables

Table 2.1: Means of raw private and shared evaluations across conditions for each of the three essay evaluation criteria, as well as for overall evaluations. Numbers in parentheses signify standard deviations. Note: Overall evaluations are on a 0 to 100 scale, quality is on a 0 to 100 scale, attributes is on a 1 to 5 scale, and recommendation is on a 1 to 7 scale.

	Overall		Quality		Attributes		Recommendation	
	Private	Shared	Private	Shared	Private	Shared	Private	Shared
Compassion	24.94 (14.18)	30.31 (16.73)	32.71 (20.01)	37.54 (21.21)	2.29 (0.61)	2.45 (0.70)	1.60 (0.97)	2.03 (1.16)
Neutral	27.00 (15.29)	29.09 (16.03)	34.97 (20.67)	36.17 (20.73)	2.33 (0.65)	2.38 (0.70)	1.76 (1.19)	2.00 (1.22)

Table 2.2: Means of raw private and shared evaluations for those high and low in trait compassion for each of the three essay evaluation criteria, as well as for overall evaluations. Numbers in parentheses signify standard deviations. Note: All scores are on 0 to 100 scales. High and low compassion were defined as greater than 1 standard deviation above and below the mean of trait compassion, respectively.

	Overall		Quality		Attributes		Recommendation	
	Private	Shared	Private	Shared	Private	Shared	Private	Shared
High Compassion	44.14 (24.79)	50.13 (25.21)	48.96 (23.83)	55.30 (26.59)	49.33 (22.02)	51.95 (22.43)	34.13 (32.51)	43.13 (31.50)
Low Compassion	36.54 (18.64)	40.75 (19.15)	45.06 (21.82)	45.13 (21.24)	43.18 (18.81)	48.63 (19.78)	21.38 (21.17)	28.50 (23.49)

References

Ashford, S. J., & Cummings, L. L. (1983). Feedback as an individual resource: Personal strategies of creating information. *Organizational Behavior and Human Performance*, *32*(3), 370-398.

Barsade, S. G. (2002). The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly*, *47*(4), 644-675.

Batson, C. D. (1991). *The altruism question: Toward a social-psychological answer*. Hillsdale, NJ: Erlbaum.

Batson, C. D., Duncan, B. D., Ackerman, P., Buckley, T., & Birch, K. (1981). Is empathic emotion a source of altruistic motivation?. *Journal of Personality and Social Psychology*, *40*(2), 290-302.

Batson, C. D., Sager, K., Garst, E., Kang, M., Rubchinsky, K., & Dawson, K. (1997). Is empathy-induced helping due to self–other merging?. *Journal of Personality and Social Psychology*, *73*(3), 495-509.

Batson, C. D., & Shaw, L. (1991). Evidence for altruism: Toward a pluralism of prosocial motives. *Psychological Inquiry*, 2, 107-122.

Black, G. (1993). *Genocide in Iraq: the Anfal campaign against the Kurds*. Human Rights Watch.

Boles, T. L., Croson, R. T., & Murnighan, J. K. (2000). Deception and retribution in repeated ultimatum bargaining. *Organizational Behavior and Human Decision Processes*, *83*(2), 235-259.

Burgoon, J. K., & Buller, D. B. (1994). Interpersonal deception: III. Effects of deceit on perceived communication and nonverbal behavior dynamics. *Journal of Nonverbal Behavior*, *18*(2), 155-184.

Broomfield, K. A., Robinson, E. J., & Robinson, W. P. (2002). Children's understanding about white lies. *British Journal of Developmental Psychology*, 20(1), 47-65.

Brummelman, E., Thomaes, S., Orobio, D. C. B., Overbeek, G., & Bushman, B. J. (2014). "That's not just beautiful--that's incredibly beautiful!": the adverse impact of inflated praise on children with low self-esteem. *Psychological Science*, *25*(3), 728-735.

Cameron, C. D., & Payne, B. K. (2011). Escaping affect: How motivated emotion regulation creates insensitivity to mass suffering. *Journal of Personality and Social Psychology, 100*, 1-15.

Cameron, C. D., & Payne, B. K. (2012). The cost of callousness: Regulating compassion influences the moral self-concept. *Psychological Science*,*23*(3), 225-229.

Cialdini, R. B., Brown, S. L., Lewis, B. P., Luce, C., & Neuberg, S. L. (1997). Reinterpreting the empathy–altruism relationship: When one into one equals oneness. *Journal of Personality and Social Psychology*, *73*(3), 481-494.

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate behavioral research*, *34*(3), 315-346.

Coke, J. S., Batson, C. D., & Mcdavis, K. (1978). Empathic mediation of helping. *Journal of Personality and Social Psychology*, *36*(7), 752-766.

Condon, P., & DeSteno, D. (2011). Compassion for one reduces punishment for another. *Journal of Experimental Social Psychology*, *47*(3), 698-701.

Côté, S., Kraus, M. W., Cheng, B. H., Oveis, C., Van der Löwe, I., Lian, H., & Keltner, D. (2011). Social power facilitates the effect of prosocial orientation on empathic accuracy. *Journal of Personality and Social Psychology*, 101(2), 217-232.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1), 113-126.

Decety, J. (2015). The neural pathways, development and functions of empathy. *Current Opinion in Behavioral Sciences*, *3*, 1-6.

Decety, J., & Cowell, J. M. (2014). Friends or foes: Is empathy necessary for moral behavior? *Perspectives on Psychological Science*, *9*(5), 525-537.

DePaulo, B. M., & Kashy, D. A. (1998). Everyday lies in close and casual relationships. *Journal of Personality and Social Psychology*, *74*(1), 63-79.

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology*, *70*(5), 979-995.

DeSteno, D. (2009). Social emotions and intertemporal choice "Hot" mechanisms for building social and economic capital. *Current Directions in Psychological Science*, *18*(5), 280-284.

DeSteno, D., Li, Y., Dickens, L., & Lerner, J. S. (2014). Gratitude: a tool for reducing economic impatience. *Psychological Science*, *25*(6), 1262-1267.

Dickens, L., & DeSteno, D. (2016). The grateful are patient: Heightened daily gratitude is associated with attenuated temporal discounting. *Emotion*, *16*(4), 421-425.

Dimberg, U., Thunberg, M., & Elmehed, K. (2000). Unconscious facial reactions to emotional facial expressions. *Psychological Science*, *11*(1), 86-89.

Eisenberg, N. (1991). Meta-analytic contributions to the literature on prosocial behavior. *Personality and Social Psychology Bulletin*, *17*(3), 273-282.

Eisenberg, N. (2002). Empathy-related emotional responses, altruism, and their socialization. *Visions of Compassion: Western Scientists and Tibetan Buddhists Examine Human Nature*, *135*, 131-164.

Eisenberg, N., & Eggum, N. D. (2009). Empathic responding: Sympathy and personal distress. *The Social Neuroscience of Empathy*, *6*, 71-83.

Eisenberg, N., & Fabes, R. A. (1990). Empathy: Conceptualization, measurement, and relation to prosocial behavior. *Motivation and Emotion*, *14*(2), 131-149.

Eisenberg, N., & Miller, P. A. (1987). The relation of empathy to prosocial and related behaviors. *Psychological Bulletin*, *101*(1), 91-119.

Eisenberg, N., Fabes, R. A., Miller, P. A., Fultz, J., Shell, R., Mathy, R. M., & Reno, R. R. (1989). Relation of sympathy and personal distress to prosocial behavior: a multimethod study. *Journal of Personality and Social Psychology*,*57*(1), 55-66.

Ellis, S., Mendel, R., & Aloni-Zohar, M. (2009). The effect of accuracy of performance evaluation on learning from experience: The moderating role of after-event reviews. *Journal of Applied Social Psychology*, *39*(3), 541-563.

Erat, S., & Gneezy, U. (2012). White lies. *Management Science*, *58*(4), 723-733.

Emerson, R. W. (1888). *Select Writings of Ralph Waldo Emerson* (Vol. 33). W. Scott.

Galinsky, A. D., Maddux, W. W., Gilin, D., & White, J. B. (2008). Why it pays to get inside the head of your opponent: The differential effects of perspective taking and empathy in negotiations. *Psychological Science*, *19*(4), 378-384.

Gino, F., Ayal, S., & Ariely, D. (2013). Self-serving altruism? The lure of unethical actions that benefit others. *Journal of Economic Behavior & Organization*, *93*, 285-292.

Gino, F., Norton, M. I., & Ariely, D. (2010). The counterfeit self: The deceptive costs of faking it. *Psychological Science*, 21(5), 712-720.

Gino, F., & Pierce, L. (2009). Dishonesty in the name of equity. *Psychological Science*, *20*(9), 1153-1160.

Graham, J., Meindl, P., Koleva, S., Iyer, R., & Johnson, K. M. (2015). When values and behavior conflict: Moral pluralism and intrapersonal moral hypocrisy. *Social and Personality Psychology Compass*, 9(3), 158-170.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366-385.

Goetz, J. L., Keltner, D., & Simon-Thomas, E. (2010). Compassion: An evolutionary analysis and empirical review. *Psychological Bulletin*, *136*(3), 351-374.

Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of Affective Sciences* (pp. 852-870). Oxford: Oxford University Press.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, *20*(1), 98-116.

Han, S., Lerner, J. S., & Keltner, D. (2007). Feelings and consumer decision making: The appraisal-tendency framework. *Journal of Consumer Psychology*, 17(3), 158-168.

Haselton, M. G., Buss, D. M., Oubaid, V., & Angleitner, A. (2005). Sex, lies, and strategic interference: The psychology of deception between the sexes. *Personality and Social Psychology Bulletin*, *31*(1), 3-23.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). Emotional contagion. *Current Directions in Psychological Science*, *2*(3), 96-100.

Hertenstein, M. J., Keltner, D., App, B., Bulleit, B. A., & Jaskolka, A. R. (2006). Touch communicates distinct emotions. *Emotion*, *6*(3), 528-533.

Hess, U., & Blairy, S. (2001). Facial mimicry and emotional contagion to dynamic emotional facial expressions and their influence on decoding accuracy. *International Journal of Psychophysiology*, *40*(2), 129-141.

Hess, U., & Fischer, A. (2013). Emotional mimicry as social regulation. *Personality and Social Psychology Review*, *17*(2), 142-157.

Hewig J., Hagemann, D., Seifert, J., Gollwitzer, M., Naumann, E., & Bartussek, D. (2005). A revised film set for the study of basic emotions. *Cognition & Emotion*, *19*(7), 1095-1109.

Heyman, G. D., Luu, D. H., & Lee, K. (2009). Parenting by lying. *Journal of Moral Education*, *38*(3), 353-369.

Hillman, L. W., Schwandt, D. R., & Bartz, D. E. (1990). Enhancing staff members' performance through feedback and coaching. *Journal of Management Development*, *9*(3), 20-27.

Hirsh, J. B., Guindon, A., Morisano, D., & Peterson, J. B. (2010). Positive mood effects on delay discounting. *Emotion*, *10*(5), 717-721.

Horberg, E. J., Oveis, C., & Keltner, D. (2011). Emotions as moral amplifiers: An appraisal tendency approach to the influences of distinct emotions upon moral judgment. *Emotion Review*, *3*(3), 237-244.

Huck, S. W., & McLean, R. A. (1975). Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin*, *82*(4), 511-518.

Hyland, F. (1998). The impact of teacher written feedback on individual writers. *Journal of Second Language Writing*, *7*(3), 255-286.

Ickes, W. (1993). Empathic accuracy. *Journal of Personality*, *61*(4), 587-610.

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, *64*(4), 349-371.

Immordino-Yang, M. H., McColl, A., Damasio, H., & Damasio, A. (2009). Neural correlates of admiration and compassion. *Proceedings of the National Academy of Sciences*, *106*(19), 8021-8026.

Jampol, L. E., & Zayas, V. (2016). The dark side of white lies: Underperforming women are told more white lies than men during performance feedback. Working Paper.

John, O. P., Donahue, E. M., & Kentle, R. (1991). The 'Big Five Inventory—version 4a and, 54. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of Personality: Theory and Research, 2(1999)*, 102-138.

Jordan, M. R., Amir, D., & Bloom, P. (2016). Are empathy and concern psychologically distinct?. *Emotion*, *16*(8), 1107-1116.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1999). International affective picture system (IAPS): Instruction manual and affective ratings. *The Center for Research in Psychophysiology, University of Florida*.

Lazarus, R.S. (1991). *Emotion and Adaptation*. New York: Oxford University Press.

Levenson, R. W., & Ruef, A. M. (1992). Empathy: A physiological substrate. *Journal of Personality and Social Psychology*, *63*(2), 234-246.

Levine, E. E., & Schweitzer, M. E. (2014). Are liars ethical? On the tension between benevolence and honesty. *Journal of Experimental Social Psychology*, *53*, 107-117.

Levine, E. E., & Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, *126*, 88-106.

Locke, E. A., & Latham, G. P. (1990). Work motivation and satisfaction: Light at the end of the tunnel. *Psychological Science*, *1*(4), 240-246.

Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes*, *65*(3), 272-292.

Loewenstein, G., & Small, D. A. (2007). The Scarecrow and the Tin Man: The Vicissitudes of Human Sympathy and Caring. *Review of General Psychology*,*11*(2), 112-126.

Lupoli, M.J., Levine, E.E., Greenberg, A.E. (2017). Paternalistic Lies. Working Paper.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633-644.

Neumann, R., & Strack, F. (2000). " Mood contagion": the automatic transfer of mood between persons. *Journal of Personality and Social Psychology*, *79*(2), 211-223.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know. *Psychological Review*, *84*(3), 231-259.

Nussbaum, M. (1996). Compassion: The basic social emotion. *Social Philosophy and Policy*, *13*(01), 27-58.

McCornack, S. A., & Levine, T. R. (1990). When lies are uncovered: Emotional and relational outcomes of discovered deception. *Communications Monographs*, *57*(2), 119-138.

Omoto A. M., Malsch, A. M., Barraza J. A. (2009). Compassionate acts: Motivations for and correlates of volunteerism among older adults. In B. Fehr B, S. Sprecher S, L.G. Underwood (Eds.), *The Science of Compassionate Love: Theory, Research, and Applications* (pp. 257–282). Malden, MA: Wiley-Blackwell.

Oveis, C., Cohen, A. B., Gruber, J., Shiota, M. N., Haidt, J., & Keltner, D. (2009). Resting respiratory sinus arrhythmia is associated with tonic positive emotionality. *Emotion*, *9*(2), 265-270.

Oveis, C., Horberg, E. J., & Keltner, D. (2010). Compassion, pride, and social intuitions of self-other similarity. *Journal of Personality and Social Psychology*, *98*(4), 618-630.

Preston, S. D., & De Waal, F. B. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences*, *25*(01), 1-20.

Rosenberg, E. L. (1998). Levels of analysis and the organization of affect. *Review of General Psychology*, *2*(3), 247-270.

Rudolph, U., Roesch, S., Greitemeyer, T., & Weiner, B. (2004). A meta-analytic review of help giving and aggression from an attributional perspective: Contributions to a general theory of motivation. *Cognition and Emotion*, *18*(6), 815-848.

Rynes, S. L., Bartunek, J. M., Dutton, J. E., & Margolis, J. D. (2012). Care and compassion through an organizational lens: Opening up new possibilities. *Academy of Management Review*, *37*(4), 503-523.

Saslow, L. R., Willer, R., Feinberg, M., Piff, P. K., Clark, K., Keltner, D., & Saturn, S. R. (2013). My brother's keeper? Compassion predicts generosity more among less religious individuals. *Social Psychological and Personality Science*, *4*(1), 31-38.

Shamay-Tsoory, S. G., Aharon-Peretz, J., & Perry, D. (2009). Two systems for empathy: a double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. *Brain*, *132*(3), 617-627.

Shweder, R., Much, N., Mahapatra, M., & Park, L. (1997). Divinity and the "big three" explanations of suffering. *Morality and Health*, *119*, 119-169.

Schweitzer, M. E., Hershey, J. C., & Bradlow, E. T. (2006). Promises and lies: Restoring violated trust. *Organizational Behavior and Human Decision Processes*, *101*(1), 1-19.

Shiota, M. N., Keltner, D., & John, O. P. (2006). Positive emotion dispositions differentially associated with Big Five personality and attachment style. *The Journal of Positive Psychology*, *1*(2), 61-71.

Singer, T., & Steinbeis, N. (2009). Differential Roles of Fairness-and Compassion-Based Motivations for Cooperation, Defection, and Punishment. *Annals of the New York Academy of Sciences*, *1167*(1), 41-50.

Slovic, P. (2007). "If I look at the mass I will never act": Psychic numbing and genocide. *Judgment and Decision Making*, *2*(2), 79-95.

Small, D. A., & Loewenstein, G. (2003). Helping a victim or helping the victim: Altruism and identifiability. *Journal of Risk and Uncertainty*, *26*(1), 5-16.

Small, D. A., Loewenstein, G., & Slovic, P. (2007). Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims. *Organizational Behavior and Human Decision Processes*, *102*(2), 143-153.

Stellar, J. E., Cohen, A., Oveis, C., & Keltner, D. (2015). Affective and Physiological Responses to the Suffering of Others: Compassion and Vagal Activity. *Journal of Personality and Social Psychology*, *108*(4), 572-585.

Stellar, J. E., Manzo, V. M., Kraus, M. W., & Keltner, D. (2012). Class and compassion: socioeconomic factors predict responses to suffering. *Emotion*, *12*(3), 449-459.

Stellar, J., Feinberg, M., & Keltner, D. (2014). When the selfish suffer: evidence for selective prosocial emotional and physiological responses to suffering egoists. *Evolution and Human Behavior*, *35*(2), 140-147.

Stiff, J. B., Kim, H. J., & Ramesh, C. N. (1992). Truth biases and aroused suspicion in relational deception. *Communication Research*, *19*(3), 326-345.

Stürmer, S., Snyder, M., Kropp, A., & Siem, B. (2006). Empathy-motivated helping: The moderating role of group membership. *Personality and Social Psychology Bulletin*, *32*(7), 943-956.

Talwar, V., Gordon, H. M., & Lee, K. (2007). Lying in the elementary school years: verbal deception and its relation to second-order belief understanding. *Developmental Psychology*, *43*(3), 804-810.

Tenbrunsel, A. E., & Messick, D. M. (2004). Ethical fading: The role of self-deception in unethical behavior. *Social Justice Research*, *17*(2), 223-236.

Tyler, J. M., Feldman, R. S., & Reichert, A. (2006). The price of deceptive behavior: Disliking and lying to people who lie to us. *Journal of Experimental Social Psychology*, *42*(1), 69-77.

Valdesolo, P., & DeSteno, D. (2011a). Synchrony and the social tuning of compassion. *Emotion*, *11*(2), 262.

Valdesolo, P., & DeSteno, D. (2011b). The virtue in vice: Short-sightedness in the study of moral emotions. *Emotion Review*, *3*(3), 276-277.

Van Boven, L., & Loewenstein, G. (2003). Social projection of transient drive states. *Personality and Social Psychology Bulletin*, *29*(9), 1159-1168.

Van Kleef, G. A., Oveis, C., van der Löwe, I., LuoKogan, A., Goetz, J., & Keltner, D. (2008). Power, distress, and compassion turning a blind eye to the suffering of others. *Psychological Science*, *19*(12), 1315-1322.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063-1070.

Wiltermuth, S. S. (2011). Cheating more when the spoils are split. *Organizational Behavior and Human Decision Processes*, 115(2), 157-168.

Wondra, J. D., & Ellsworth, P. C. (2015). An appraisal theory of empathy and other vicarious emotional experiences. *Psychological Review*, *122*(3), 411-428.

Wispé, L. (1986). The distinction between sympathy and empathy: To call forth a concept, a word is needed. *Journal of Personality and Social Psychology*, *50*(2), 314-321.

Zaki, J. (2014). Empathy: a motivated account. *Psychological Bulletin*, *140*(6), 1608-1647.

Chapter 3

Paternalistic Lies

Matthew James Lupoli[1], Emma Edelman Levine[2], Adam Eric Greenberg[3]

University of California, San Diego[1]
University of Chicago[2]
University of California, Los Angeles[3]

Correspondence concerning this article should be addressed to: Matthew J. Lupoli, Rady School of Management, University of California, San Diego, Wells Fargo Hall #4W124, San Diego, CA, 92093-0553. Phone: (917) 373-2971, Email: Matthew.lupoli@rady.ucsd.edu

Abstract

Many lies that are intended to help others require the deceiver to make assumptions about whether lying serves others' best interests. In other words, lying often involves a paternalistic motive. Across seven studies (N = 2,260), we show that although targets appreciate lies that yield unequivocal benefits relative to honesty, they penalize paternalistic lies. We identify three mechanisms behind the harmful effects of paternalistic lies, finding that targets believe that paternalistic liars (a) do not have benevolent intentions, (b) are violating their autonomy by lying, and (c) are inaccurately predicting their preferences. Importantly, targets' aversion towards paternalistic lies persists even when targets receive their preferred outcome as a result of a lie. Additionally, deceivers can mitigate some, but not all, of the harmful effects of paternalistic lies by directly communicating their good intentions. These results contribute to our understanding of deception and paternalistic policies.

Preface

Imagine a man who has been drinking gets behind the wheel and accidentally crashes into a tree, damaging his vehicle but injuring no one. Now, imagine that instead of hitting a tree, the man hits and kills a young child. Should these crimes receive the same punishment? Consider a third scenario: the man sees a child in the road, actively *tries* to hit the child, and ends up killing him. Should this be punished in the same way as if hitting the child were accidental?

Most people would answer "no" to both of these questions. What this illustrates is that when making moral judgments, people take into consideration both intentions and consequences of actions (Cushman, 2008; 2013; Gino, Shu, & Bazerman, 2010; Greene et al., 2009; Miller, Hannikainen, & Cushman, 2014). In Chapter 3, I explore the interplay between intentions and consequences in the context of moral judgments of deception. More specifically, I investigate the moral judgments people make of deceivers when they learn that they have been the target of a lie that was intended to benefit them. As it turns out, neither intentions nor consequences are alone sufficient to explain how people respond to these lies, but they are both necessary.

In Chapter 1, I argued why prosocial lies are better suited to be defined by intentions rather than consequences. One of the reasons I cited is that consequences are uncertain—we don't always know how prosocial lies will affect others. Yet, sometimes we find out after the lie is told. When people discover that they have been lied to, this creates a moment where the effects of a lie materialize; people learn not only that they were told a falsehood, but they also may learn the truth from which they were previously deprived. Amongst the many thoughts that might arise upon acquiring insight into both the truth and the lie and all of their implications, people may draw conclusions about which makes them better off. That is, an individual might consider, "*should* this person have lied to me?" In Chapter 3, I demonstrate that intentions alone are not

sufficient to determine how people respond when they learn that they have been the target of a lie that was intended to help them. One must also consider the consequences—that is, the benefits and drawbacks associated with both honesty and dishonesty, in the eyes of the target. It is through this consideration that the construct of *paternalistic lies* was born.

Paternalistic lies are those that are intended to benefit the target, but that require the deceiver to make assumptions about the target's best interests. By this definition, paternalistic lies are a subset of prosocial lies. However, unlike the prosocial lying, this construct takes a hybrid approach to the role of intentions and consequences by incorporating both in the definition: While the intentions are prosocial, it is uncertain whether the outcome of the lie is beneficial to the target. In Chapter 3, I contrast these lies with *unequivocal prosocial lies*, or those that have unequivocal benefits for the target, and show how this type of lying can lead to strongly divergent responses amongst targets compared with paternalistic lies.

In my view, the incorporation of consequences into the definition of these constructs does not contradict my previous position on the importance of intention-based definitions of prosocial lies. Instead, it signifies the need to go beyond intentions when studying responses to prosocial lies. As discussed in Chapters 1 and 2, prosocial lies can backfire, and sometimes it is simply the revelation of the lie that constitutes that backfiring. Because lies do become uncovered in the real world, it is important to examine what happens when people find out that they have been lied to. In Chapter 3, I illustrate how both intentions and consequences of paternalistic and unequivocal prosocial lies influence responses to these lies.

This research sits amongst a growing body of work highlighting the many nuances of responses to prosocial lies (Levine, 2018; Levine & Schweitzer, 2014, 2015; Levine et al., 2018). Although people's reactions to these lies are complex, we are now beginning to understand the

systematic rules that govern them. As we continue to learn more about when and why people do and do not find prosocial lying permissible, we can hopefully improve interpersonal communication and social life in general by making fewer mistakes about what we think will help people, when those who are "helped" would disagree.

Paternalistic Lies

People often lie with the intention of benefitting others (DePaulo, Kashy, Kirkendol, Wyer, & Epstein, 1996). In many cases, however, it is not immediately obvious whether lying will, in fact, benefit the recipient of the lie (henceforth the "target"). For example, an employee may inflate impressions of a colleague's performance on a presentation because he believes honesty will cause emotional harm and demotivate the colleague. Yet this belief may not necessarily be correct. A truthful statement might be seen as more beneficial in the eyes of the colleague, and could actually motivate the colleague to learn from his shortcomings and improve his performance in the future. If this colleague were to find out that the employee lied about her performance, how might he react?

In this research, we investigate how targets respond to lie-tellers (henceforth "deceivers" or "liars") whose lies require them to make subjective judgments about the target's best interests. We label these lies as paternalistic lies. Paternalistic lies are ubiquitous and have important consequences in a variety of contexts. For example, government officials might tell paternalistic lies to citizens by concealing facts about potential security threats to avoid inciting national panic; doctors might tell paternalistic lies to patients by providing overly optimistic prognoses to patients to provide hope; and friends and romantic partners might tell paternalistic lies to each other by delivering false praise with the intention of preventing emotional harm. In all of these cases, deceivers might lie out of genuine concern for the well-being of the targets, but targets may not appreciate these lies because judgments about whether the lie is ultimately more beneficial than the truth are inherently subjective. Thus, well-intended paternalistic lies may backfire. Because paternalistic lies are prevalent and can have important effects on people's lives, it is crucial to understand how they influence interpersonal judgment and behavior.

Here, we provide the first investigation of paternalistic lies. In addition to providing

practical advice to those who might be tempted to tell paternalistic lies, we fill an important gap in existing deception research by introducing the construct of paternalistic lies, distinguishing this construct from related forms of deception, and documenting a strong distaste towards paternalistic lies and those who tell them across several dependent variables. This research also deepens our understanding of the primacy of perceived intent in moral judgment; we find that the perceived intentions of paternalistic liars play a critical role in responses to these lies.

Prosocial and Paternalistic Lies

Research investigating the consequences of deception has linked lying with a number of harmful effects. Lies have been shown to increase negative affect, damage trust, provoke revenge, harm relationships, and promote further dishonesty (Boles, Croson, & Murnighan, 2000; Croson, Boles, & Murnighan, 2003; Greenberg, 2016; Greenberg & Wagner, 2016; Schweitzer & Croson, 1999; Schweitzer, Hershey, & Bradlow, 2006; Tyler, Feldman, & Reichert, 2006). However, the majority of this work has studied the effects of *selfish lies*, or lies that benefit the deceiver, potentially at a cost to the target. Given the conflation of deception with self-interested motivations in much of the existing literature, it has been difficult to conclude whether interpersonal penalties towards deception reflect an opposition to selfish behavior or deception per se.

To shed light on this issue, scholars have recently examined the consequences of prosocial lies. People tell *prosocial lies*, or false statements made with the intention of misleading and benefitting a target (Levine & Schweitzer, 2014, 2015; Lupoli, Jampol, & Oveis, 2017), on a regular basis (DePaulo et al., 1996). Given that individuals not only consider actions, but also the intentions behind and the consequences of those actions when making moral judgments of themselves (Shalvi, Dana, Handgraaf, & De Dreu, 2011; Shalvi, Gino, Barkan, &

Ayal, 2015) and others (Cushman, 2008; 2013; Gino, Shu, & Bazerman, 2010; Greene et al., 2009; Miller, Hannikainen, & Cushman, 2014; Shu, Gino, & Bazerman, 2011), it is likely that prosocial lies are perceived differently than selfish lies.

Indeed, recent work provides evidence for this assertion. Individuals who tell prosocial lies that yield objective monetary benefits to the target are viewed as more ethical than those who tell the truth, regardless of whether the deceiver benefitted from lying (Levine & Schweitzer, 2014). Importantly, this research demonstrates that positive moral judgments of prosocial liars are driven by the perceived benevolence, rather than honesty, of the deceiver. In addition, prosocial liars are sometimes perceived to be more trustworthy: for example, Levine and Schweitzer (2015) found that individuals were more likely to pass money in a trust game to those who told a prosocial lie than those who told harmful truths. Although prosocial lies increased benevolence-based trust (the willingness to make oneself vulnerable based on beliefs about another person's good intensions, which is captured by the trust game), the authors also found that prosocial lies harmed integrity-based trust—that is, the willingness to make oneself vulnerable based on beliefs about another person's adherence to moral principles, such as honesty and truthfulness. Thus, reactions towards prosocial lies are not universally positive.

While this research has advanced our understanding of prosocial lies, it has focused on one specific type of prosocial lie: lies with objective monetary benefits. Specifically, the majority of research on prosocial lies has utilized economic games to study the decisions to lie (Erat & Gneezy, 2012), as well as reactions to lying (Levine & Schweitzer, 2014, 2015). In these studies, lying is unambiguously beneficial for the target relative to the truth because a dishonest statement from a deceiver results in a monetary gain for the target, the magnitude of which exceeds the payoff resulting from honesty. Other work has investigated prosocial lying that helps

a third party, whereby individuals cheat on a task for the monetary benefit of another individual (Gino, Ayal, & Ariely, 2013; Gino & Pierce, 2009; Wiltermuth, 2011). We conceptualize these lies as *unequivocal prosocial lies* because lying is known to both the target and the deceiver to be in the best interest of the target or third party. When liars tell unequivocal prosocial lies, targets perceive the liars' benevolent intentions to be sincere, and thus, targets react favorably to deception (Levine & Schweitzer, 2014).

However, in many cases, both the consequences and true intentions associated with prosocial lies are unclear. For example, imagine that an employee (Bob) asks a colleague (Joe) for feedback on a presentation. When Bob asks Joe how he performed, what should Joe say? One option is to provide an honest opinion, believing that Bob would prefer to hear the truth and that knowing his presentation was unsatisfactory might help him improve in the future. Alternatively, Joe could lie to Bob, believing that Bob is looking for positive reinforcement and that hearing his performance was poor would devastate him. Without knowing how the truth or a lie would affect Bob emotionally or help him in the future, Joe must rely on his assumptions about Bob's best interests when deciding whether to be truthful. This scenario illustrates that when given the opportunity to tell a prosocial lie, individuals often lack insight into others' preferences for truthfulness, as well as the negative consequences lying might have on them. Thus, this type of lie can be considered a *paternalistic lie*.

We define paternalistic lies as *lies that are intended to benefit the target, but require the deceiver to make assumptions about targets' best interests.* As such, paternalistic lies are a subset of prosocial lies (see Table 3.1). When individuals tell paternalistic lies, they are motivated by the assumption that targets are better off being lied to, even though this assumption cannot be objectively verified. Thus, the targets themselves might not agree with this assessment. In short,

while unequivocal prosocial lies are known to help the target, paternalistic lies help the target only according to the beliefs of the deceiver. By studying paternalistic lies, we build knowledge of how different types of lies influence interpersonal judgment and behavior, and gain insight into the circumstances in which targets believe versus discredit the prosocial intentions of liars.

It is important to note that although we dichotomize the distinction between unequivocal prosocial lies and paternalistic lies for the ease of investigation, the degree to which the deceiver has insight into the target's best interests—and thus the degree to which a lie is paternalistic—falls along a continuum. We use the terms "paternalistic lies" and "unequivocal prosocial lies" as endpoints on this continuum. We do not claim that there are lies that are unequivocally prosocial to all people in all settings. However, we do claim that there are cases in which a deceiver can be more or less confident about what benefits the target. For instance, consider the aforementioned example of Joe, who is asked to give feedback to his colleague Bob on Bob's poor presentation. If the two have an existing relationship and have already discussed Bob's preferences for blunt critiques or words of encouragement, Joe's assumptions about whether honesty or deception are in his colleague's best interests may be fairly accurate. However, if the two have no existing relationship, then his assumptions will be less informed. Without explicit knowledge of how a lie will affect the target and the target's preferences for lying itself, lying with prosocial intent always requires some assumption regarding the target's interests. That said, if a deceiver is able to gain insight into the target's preferences (e.g., through discussion or past experience), then the deceiver is no longer required to rely on as many assumptions. Thus, lies are distinguishable with respect to how paternalistic they are.

The distinction between paternalistic lies and unequivocal prosocial lies is not merely theoretical, but one that lay people recognize as well. In a pilot study ($N = 90$), we asked

participants to generate an example of one circumstance in which someone lied with the intention of helping or protecting someone else. We then asked participants to categorize their example as either a paternalistic lie or an unequivocal prosocial lie. A total of 36% of participants indicated that "the liar assumed that lying was in the person's best interests without knowing for certain" (i.e., told a paternalistic lie), rather than "the liar knew for certain that lying was in the person's best interests" (i.e., told an unequivocal prosocial lie). For example, one participant gave the example of a person giving overly positive feedback of another's appearance, and offered the following explanation: "It might actually be in the other person's best interest to tell them they don't look good, if this would cause them to change something about their appearance that would lead to better treatment and higher self-esteem." We also asked participants to rate how often they have been the target of both types of lies and found that participants believe they are told unequivocal prosocial lies and paternalistic lies with equal frequency.[15] Together, these results suggest that (a) people recognize that some lies are paternalistic, according to our definition, (b) people perceive being the target of paternalistic lies as often as being the target of unequivocal prosocial lies, and (c) people distinguish between

[15]In addition, participants read three vignettes depicting paternalistic lies (adapted from the vignettes used in Study 7) and were asked for each vignette, "to what extent is this lie paternalistic? By paternalistic, we mean limiting the freedom or autonomy of the person who has been lied to, in the presumed best interest of that person" (1 = *not at all*, 7 = *very much so*). Collapsing across vignettes, participants rated the lies as significantly paternalistic ($M = 5.20$, $SD = 1.31$; $t(89) = 8.73$, $p < .001$, one sample t-test against a mean of 4). We provide additional details on this pilot study in our online supplementary materials.

paternalistic and unequivocally prosocial lies. Given that paternalistic lies are common, consequential, and viewed as distinct from unequivocal prosocial lies, it is important to understand their consequences.

Perceptions of Paternalistic Lies

Our central thesis is that those who tell paternalistic lies are judged to be less moral than those who are honest. To explain this prediction, we draw on three streams of research: research on procedural justice (e.g., Brockner et al., 1994; Tyler, Degoey, & Smith, 1996), research on the primacy of perceived intentions in moral judgments (e.g., Cushman, 2008; Greene et al., 2009), and research on reactance and the importance of individual autonomy (e.g., Brehm, 1966).

The procedural justice literature suggests that paternalistic lies, unlike unequivocal prosocial lies, will be viewed harshly. A robust finding in the justice literature is that the desirability of outcomes and the perceived fairness by which those outcomes are obtained interact to influence responses to outcomes (for a review, see Brockner & Wiesenfeld, 1996). Specifically, if people view an outcome as desirable, they will respond favorably regardless of the fairness of the process that yielded the outcome. However, if the outcome is undesirable, their response hinges on the perceived fairness of the process that yielded the outcome: people will respond more favorably if the process seemed fair and less favorably if the process seemed unfair. For example, one study found that organizational commitment was relatively unaffected by the perceived fairness of procedures when satisfaction with job outcomes (e.g., compensation) was high; however, when satisfaction with outcomes was low, organizational commitment was strongly influenced by procedural fairness (McFarlin & Sweeny, 1992). This pattern of results has been observed across a wide range of dependent variables, including job performance, job satisfaction, and trust in management, in both organizational and laboratory contexts (Brockner

& Wiesenfeld, 1996).

Surprisingly, no existing work has applied this lens to the study of deception. We build on procedural justice research to explain why individuals have positive reactions towards unequivocal prosocial lies, but may have negative reactions towards paternalistic lies. By definition, unequivocal prosocial lies result in outcomes that are objectively desirable (compared to the outcomes associated with honesty). Thus, in line with the procedural fairness/outcome desirability interaction, individuals are likely to respond favorably to these lies despite potentially objecting to the process (i.e., deception) in general. Indeed, this notion is consistent with past findings on positive perceptions of prosocial lies (Levine & Schweitzer, 2014, 2015). Paternalistic lies, however, result in outcomes that are not objectively desirable (compared to the outcomes associated with honesty). Thus, when people are targets of paternalistic lies, they are likely to shift their focus towards the process by which outcomes are obtained (i.e., deception or honesty). Because honesty is generally perceived to be more moral than deception (Graham, Meindl, Koleva, Iyer, & Johnson, 2015)—particularly in the absence of clear benevolent motives for deception (Levine & Schweitzer, 2014)—we expect that those who tell paternalistic lies will be judged as less moral than those who tell the truth.

What specific inferences about those who tell paternalistic lies might underlie a potential decrease in perceived moral character? We hypothesize that the perceived intent of deceivers plays a key role in moral judgments, and in particular, that targets will view paternalistic deceivers as not acting with benevolent intent. Moral judgments of actions often hinge on the perceived motives of the actor (e.g., Cushman, 2008). Consistent with this notion, past work on unequivocal prosocial lies suggests that these lies are seen as moral precisely because they credibly signal benevolent intent (Levine & Schweitzer, 2014, 2015). Lies that do not signal

benevolent intent, in contrast, are deemed to be less moral than the truth (Levine & Schweitzer, 2014). We propose that paternalistic lies signal a lack of benevolent intent for two reasons. First, the subjective nature of the benefits afforded by paternalistic lies may obscure the good intentions of deceivers. People's ability to take the perspective of others and understand the emotions, beliefs, and motivations that drive them is notably limited (e.g., Epley, Keysar, Van Boven, & Gilovich, 2004; Gilbert & Malone, 1995; Van Boven & Loewenstein, 2003). Thus, if a target thinks that he may have been better off or equally well off receiving the truth, he may incorrectly think that the deceiver with good intentions was also aware of this belief. Furthermore, personal experience with the harmful effects of selfish lies (Boles et al., 2000; Croson et al., 2003; Greenberg, Smeets, & Zhurakhovska, 2015; Greenberg & Wagner, 2016; Schweitzer & Croson, 1999; Schweitzer et al., 2006; Tyler et al., 2006) may have spillover effects on responses to prosocially motivated lies. As a result, individuals might generally be skeptical of deceivers' prosocial intentions, unless the benefits of lying over honesty are clear and unequivocal.

In addition to hypothesizing that paternalistic lies lead targets to doubt deceivers' benevolent motivation, we predict that paternalistic lies are perceived to violate targets' autonomy. Autonomy has been defined as the perceived internal locus of causality (deCharms, 1968; Ryan & Deci, 2000), or a sense that one's actions "emanate from oneself and are one's own" (Deci & Ryan, 1987). Autonomy has been found to thrive when individuals experience choice, when others acknowledge their feelings, and when individuals have the ability to take self-directed actions (Deci & Ryan, 1985). In contrast, autonomy can be diminished by deadlines, directives, pressured evaluations, and imposed goals (Ryan & Deci, 2000).

One reason why paternalistic lies might be seen as a violation of one's autonomy is that people feel they have a right to know the truth, and that acts of dishonesty impinge upon this right. Similarly, lying might be seen as an attempt to control someone else's view of the world, imposing a framework on targets that deceivers deem superior to the reality shaped by the truth. Indeed, philosophers ranging from Kant (1785) to Bok (1978) have opposed deception on these same grounds. Paternalistic lies might also threaten targets' autonomy because these lies, by definition, result in an outcome that the target may not have chosen for himself. Thus, paternalistic lying is likely to be perceived as an attempt to influence or coerce the target. Unequivocal prosocial lies, in contrast, generate an outcome that is known to be in the target's best interest. In other words, unequivocal prosocial lying is the course of action that the target would have chosen for himself. Thus, it is less likely that unequivocal prosocial lies would be perceived as autonomy violations. Given the importance of autonomy to moral judgment (Rozin, Lowery, Imada, & Haidt, 1999; Shweder, Much, Mahapatra, & Park, 1997), we predict that the perception that paternalistic lies violates one's autonomy will further contribute to judgments of deceivers' immorality.

If perceived autonomy violations underlie moral judgments of paternalistic liars, this finding would suggest that paternalistic lies elicit reactance. Reactance is a psychological state that arises when individuals feel that their freedom or autonomy is being eliminated or threatened by another (Brehm, 1966). This state can manifest as the derogation of the agent restricting the freedom (Miron & Brehm, 2006). Judging those who tell paternalistic lies to be less moral than those who are honest is one way in which targets might derogate deceivers who are perceived to be violating their autonomy. However, another indicator of reactance that could result from paternalistic lies is a decrease in attractiveness the outcome resulting from the lie (Brehm, Stires,

Sensenig, & Shaban, 1966). For example, recommendations by experts that contradict consumers' initial impressions cause consumers to oppose the recommendations more intensely because they experience a state of reactance (Fitzsimons & Lehman, 2004). Similarly, if paternalistic lies elicit reactance, targets' preferences may shift as a result of being lied to. Specifically, targets may dislike outcomes associated with lying, even if they would have liked the same outcome had it been associated with honesty. As a result, targets may feel that paternalistic liars are incorrectly predicting their preferences. If targets believe that deceivers made a wrong decision on their behalf—a decision that is potentially seen as immoral—it is possible that this could result in a halo effect (e.g., Nisbett & Wilson, 1977) whereby the deceiver is also viewed as immoral. Thus, it is possible that perceptions that deceivers inaccurately predicted one's preferences may influence moral judgments of paternalistic lies.

In summary, we consider three potential processes that may underlie moral judgments of those who tell paternalistic lies: perceptions that (a) paternalistic liars are not motivated by benevolent intent, (b) that paternalistic lies violate one's autonomy, and (c) that paternalistic liars are inaccurately predicting targets' preferences. We expect that these processes can operate in tandem, but that each independently influences moral judgments.

Overview of Studies

In seven experiments, we provide the first investigation of paternalistic lies by examining how individuals judge paternalistic lies and those who tell them. We focus primarily on moral judgments of paternalistic deceivers (Studies 1-3, 5-7). We also measured positive affect (Studies 1-3, 5-6) to assess psychological responses to paternalistic lies, in addition to social judgments of deceivers. In Studies 1-5, we examined judgments of paternalistic lies in a well-controlled economic game in which the consequences of lying (relative to truth-telling) for the target were

directly manipulated. In Study 1, we explored how both paternalistic lies and unequivocal prosocial lies influenced moral judgments and emotional responses. In Study 2, we conceptually replicated Study 1 with a larger sample size and eliminated a potential confound. In Study 3, we examined the mechanisms underlying the effect of paternalistic lies on moral judgments. In Study 4, we moved beyond moral judgments and affect by exploring how paternalistic lies alter preferences for the outcomes associated with lying and honesty. In Study 5, we documented the robustness of our results by (a) using a behavioral measure to capture targets' distaste for paternalistic lies—that is, the degree to which targets punish their deceivers—and (b) testing whether the distaste for paternalistic deception persists even when deceivers communicate their benevolent intentions. We also provided further evidence for the underlying mechanisms identified in Study 3. In Study 6, we assessed external validity of these results by examining judgments of paternalistic and unequivocal prosocial lies in several realistic vignettes. In these vignettes, we again manipulated whether the deceiver communicated benevolent intentions to the target. In Study 7, we use a vignette design similar to that of Study 6 to directly manipulate the deceiver's benevolent intent, rather than the deceiver's *claimed* benevolent intent, to obtain causal evidence for a hypothesized mechanism underlying moral judgments of paternalistic liars.

Deception Game

A large body of research demonstrates the capacity for economic games to teach us about decision-making in real-world dilemmas and social interactions (e.g., Fehr & Fishbacher, 2003; Halevy & Chou, 2014; Halevy & Halali, 2015; Murnighan & Wang, 2016; Zhong, 2011). Games have several advantages, including clean experimental control over endogenous and exogenous factors, ease of comparison across experimental designs and results (Ostrom, Gardner, & Walker, 1994), and unambiguously defined actions and consequences for players (Rapoport,

1973). Given these advantages, deception has often been studied using variations of an economic game called the sender-receiver game (Erat & Gneezy, 2012; Gneezy, 2005; Gneezy, Rockenback & Serra-Garcia, 2013; Gunia, Wang, Huang, Wang, & Murnighan, 2012; Levine & Schweitzer, 2014, 2015; Zhong, 2011). Although different types of lies have been operationalized using this game (e.g., altruistic lies that benefit others at a cost to oneself; Erat & Gneezy, 2012), no work that we are aware of has used the game to explore paternalistic lies. Thus, in Studies 1-5, we adapted a version of the sender-receiver game to study paternalistic lies, hereafter referred to as the Deception Game.

In this game, all participants learned that they had been assigned to the role of "Receiver," and that they were paired with an anonymous "Sender." In actuality, there was no Sender; the Sender's role was simulated by the computer. Participants were told that the computer had simulated a fair coin flip, and that only the Sender knew the actual outcome of the coin flip. They were informed that after learning the outcome of the flip, the Sender sent one of two messages to the Receiver (participants): "The coin landed on HEADS" or "The coin landed on TAILS." Participants were then told that after receiving the Sender's message, they would choose "heads" or "tails," and what they earned would be based on whether their choice corresponded to the actual outcome of the coin flip. Importantly, both the Sender and the Receiver knew that only the Sender was informed about the potential payoffs associated with the Receiver's choice.[16]

[16]After reading the instructions, participants completed a comprehension check to ensure that they understood the instructions. If they answered either of the comprehension check questions incorrectly, they were given the exercise instructions again, followed by a second comprehension

Next, participants were randomly assigned to receive one of the two possible messages ostensibly from the Sender (i.e., "The coin landed on HEADS/TAILS"). After viewing the message, participants were asked to choose either "heads" or "tails."

Once they made their choice, participants were told that they would learn about the private information that was available to the Sender—that is, the possible payoffs and full instructions the Sender received. We then revealed the Sender's information to participants. Specifically, participants learned three new pieces of information. First, they learned that the outcome of the coin flip was heads. This constituted our between-subjects manipulation of (dis)honesty. Those who were told by the Sender that the coin landed on heads received the truth, while those who were told that the coin landed on tails were deceived.

Second, Receivers learned that Senders were told, "previous studies have found that almost all Receivers choose the outcome that the Sender indicates in his/her message." We included this statement because we wanted participants to believe the Sender would expect them to follow the message. This was intended to reduce noise in participants' perceptions of whether the Sender lied, since otherwise participants could have thought that the Sender would expect them to not follow her message (Sutter, 2009).

Third, participants learned about the payment structure that the Sender faced. According to the Sender's instructions, if the Receiver chose correctly (i.e., her choice corresponded to the actual outcome of the coin flip), the Receiver would be paid according to Option A. If the Receiver chose incorrectly (i.e., her choice did not correspond to the actual coin flip outcome),

---

check. If they failed the second comprehension check, they were unable to continue with the experiment.

the Receiver would be paid according to Option B. As such, Senders had faced the choice of sending an honest message, which would likely result in the Receiver getting Option A, or sending a dishonest message, which would likely result in the Receiver obtaining Option B. Importantly, the Sender's own incentives were not tied to either Option A or Option B. Thus, the Sender was simply making a decision that would affect the Receiver, not herself.

In all studies employing the deception game (Studies 1-5), Option A and Option B were pretested to be equally desirable in the aggregate, but involved some tradeoff that could be perceived differently at the individual level. For example, in Study 1, one Option was a low-risk, low-reward gamble, while the other Option was a higher-risk, higher-reward gamble. Structuring the game such that the outcomes were equally desirable on average simulates conditions under which a paternalistic lie might be told; from the Sender's perspective, there is necessarily uncertainty about which outcome is in the Receiver's best interest. Lying to ensure that an individual received a low-risk, low-reward gamble may be well-intended, given that it protects the target from some risk. Yet, this lie is necessarily paternalistic because the deceiver does not know what the target's risk preferences are, and thus, must make assumptions about what the target would want. Indeed, a pretest revealed that Senders who lied in the Deception Game did so because they believed it was in the best interest of Receivers.[17] However, from the Receiver's

[17]We ran a pilot study in which all participants were assigned to the role of Sender ($N = 148$). After making the decision to send an honest or dishonest message, Senders were asked to indicate their agreement with the statement, "I chose the message I believed was in the best interest of the Receiver" (1 = *strongly disagree*, 7 = *strongly agree*). A t-test against the midpoint indicated that Senders who lied ($N = 44$) significantly agreed with this statement ($M = 5.32$, $SD = 2.18$), $t(43) =$

perspective, the Sender's motivations are intentionally ambiguous because in the real world, targets often are not fully aware of deceivers' motives. Table 3.2 includes a summary of the outcomes associated with Options A and B for each study, along with an example of the type of paternalistic lies these options model.

In each study, we counterbalanced the outcomes associated with Options A and B between-subjects to ensure that our results were robust across any particular tradeoff. For example, in Study 1, half of the participants saw that Option A was the low-risk, low-reward gamble, and that Option B was the high-risk, high-reward gamble; and the other half of participants saw that Option A was the high-risk, high-reward gamble, and Option B was the low-risk, low-reward gamble. Furthermore, in all studies, the potential payoffs in the game were incentive-compatible, as one participant was randomly selected to receive the Option obtained in the game.

In all studies, we did not conduct statistical analyses prior to the completion of data collection. We report all measures and manipulations. Given that we did not have sufficient precedent to make precise estimates of effect sizes, we decided on sample size using the following heuristics: For laboratory studies (Studies 1 and 4), we aimed to obtain as many participants as possible within the lab time allotted; for online studies (Studies 2-3, 5-7), we aimed to obtain 100 participants per cell (collapsed across choice set for studies using the Deception Game). We report all measures, manipulations, and data exclusions.

Study 1

4.00, $p < .001$, suggesting that their deception was motivated by their assumptions about what benefitted the Receiver (i.e., their deception was paternalistic).

In Study 1, we investigated individuals' moral judgments of those who tell paternalistic lies. In this experiment, both honesty and dishonesty resulted in participants being entered into one of two gambles that were equally desirable on average. Using gambles with different levels of risk as outcomes captures the uncertainty often associated with paternalistic lies. For example, a mentor might lie to an employee if she thinks a low-risk, low-reward career is better for the candidate than a high-risk, high-reward career. Similarly, a doctor might lie to a patient to lead her to choose a low-risk (or high-reward) treatment.

In order to disentangle whether reactions to the Sender were due to the subjective nature of the message's consequences or reactions to deception in general, we also included conditions in which the Sender's message resulted in participants being given one or two lottery tickets for entry into the same gamble. In other words, we compared paternalistic lies (lies that require the deceiver to make assumptions about the best interests of the target) to unequivocal prosocial lies (lies that are known to the target and the deceiver to be in the best interest of the target).

Procedure and Materials

We recruited 200 adults from a city in the northeastern United States to participate in a study in exchange for a $10 show-up fee. Eight participants failed a comprehension check at the start of the experiment and were automatically eliminated from the study. We thus report the results from 192 participants (59.9% female; M_{age} = 20) who passed the comprehension checks and completed the entire study.

Participants were randomly assigned to one of eight experimental conditions in a 2(Deception: honesty vs. lying) x 2(Lie type: paternalistic lie vs. unequivocal prosocial lie) x 2(Choice Set: choice set 1 vs. choice set 2) between-subjects design. In the paternalistic lie

conditions, the benefit associated with lying was subjective. In the unequivocal prosocial lie conditions, deception was unambiguously prosocial (i.e., it made the target strictly better off).

Participants engaged in the Deception Game as previously described. After learning the rules of the game and receiving the randomly assigned message from the Sender, we revealed the Sender's private information. Participants learned that the Sender had either been honest or dishonest. We also revealed the payoffs associated with Options A and B, which were lotteries in which Receivers would be entered. Here, we manipulated whether or not the Sender's lie could yield an objectively beneficial payout. In the paternalistic lie conditions, Options A and B were associated with a 50% chance of winning $1 and a 50% chance of winning $0, versus a 25% chance to win $2.25 and a 75% chance of winning $0. These gambles were rated as equally preferable ($p > .40$) in a pilot study with a non-overlapping sample ($N = 46$). As mentioned, we counterbalanced the outcomes associated with Options A and B between-subjects so that there were two different choice sets. That is, in the choice set 1/paternalistic lie condition, Option A resulted in the Receiver getting 1 lottery ticket for the 50% chance of $1/50% chance of $0 lottery, while Option B yielded 1 lottery ticket for the 25% chance of $2.25/75% chance of $0 lottery. In the choice set 2/paternalistic lie condition, the lotteries associated with Options A and B were reversed.

In the unequivocal prosocial lie conditions, Options A and B were associated with one lottery ticket or two lottery tickets for identical gambles, respectively. As in the paternalistic lie conditions, we counterbalanced the types of gambles associated with Options A and B so that there were two different choice sets. In the choice set 1/unequivocal prosocial lie condition, participants saw that Option A resulted in the Receiver receiving 1 lottery ticket for the 50% chance of $1/50% chance of $0 gamble, and that Option B resulted in 2 lottery tickets for this

same gamble. The other half of participants (in the choice set 2/unequivocal prosocial lie condition) saw that Option A yielded 1 ticket and Option B yielded 2 tickets for the 25% chance of $2.25/75% chance of $0 gamble. Because Option B dominates Option A in the both choice sets of the unequivocal prosocial lie conditions, sending a dishonest message would result in an outcome that was objectively better for the Receiver.

Dependent variables. After viewing the Sender's private information and the potential outcomes associated with Options A and B, participants provided ratings of the Sender's moral character by indicating their agreement (1 = *strongly disagree*, 7 = *strongly agree*) with the following statements: "I trust the Sender"; "The Sender had good intentions"; "The Sender wanted to help me"; "The Sender is a good person"; "The Sender is unethical" (reverse-scored); and "The Sender made the wrong decision for me" (reverse-scored). This last item was not included in analysis of moral judgments, as it conflates perceptions of the Sender with personal preferences. However, inclusion of this item does not alter results. The remaining items were highly reliable ($\alpha = .89$).

We also measured participants' emotional responses to the Sender's message. Participants were asked to indicate the degree to which they felt the following emotions "in response to the Sender's behavior" (1 = *not at all*, 7 = *extremely*): grateful, excited, happy, and content ($\alpha = .88$).[18]

[18]In addition, we measured negative affect using the following four items: angry, disappointed, sad, and anxious ($\alpha = .82$). Positive and negative affect loaded on separate factors. We measured both positive and negative affect in Studies 1-3, 5, and 6. However, for the sake of brevity, we

Finally, we included a three-item manipulation check to ensure that participants recognized the act of deception. Participants rated their agreement (1 = *strongly disagree*, 7 = *strongly agree*) with the following items: "The Sender sent an honest message" (reverse-scored); "The Sender lied about the outcome of the coin flip;" and "The Sender was deceptive" ($\alpha = .86$). Participants concluded the study by providing demographic information and answering three attention checks.

Results

For all studies, we report results collapsing across choice set for the sake of brevity. Models with choice set included as a factor are included in the Supplementary Materials.

Manipulation check. A t-test revealed that the deception manipulation was successful. Collapsing across lie type and choice set, participants in the lie condition ($M = 5.56$, $SD = 1.39$) rated the Sender as more dishonest than those in the truth condition ($M = 2.33$, $SD = 1.38$), $t(190) = 16.74$, $p < .001$, $d = 2.42$.

Moral character. A two-way ANOVA revealed a significant Deception x Lie Type interaction, $F(1, 188) = 26.15$, $p < .001$, $\eta_p^2 = .12$. Consistent with Levine and Schweitzer (2014), Senders were seen as more moral when they told an unequivocal prosocial lie ($M = 4.54$, $SD = 1.70$) than when they told the truth ($M = 3.99$, $SD = 1.10$), $t(92) = 1.88$, $p = .06$, $d = .39$. Importantly, however, this effect reversed for paternalistic lies. When lying was associated with subjective benefits, Senders were seen as more moral when they told the truth ($M = 4.94$, $SD = 1.02$) than when they lied ($M = 3.62$, $SD = 1.19$), $t(96) = 5.93$, $p < .001$, $d = 1.20$. This pattern of

---

only report positive affect. Positive and negative affect followed the inverse pattern in every study and the results for negative affect are reported in the Supplementary Materials.

results is depicted in Figure 3.1. In addition, those who told paternalistic lies (M = 3.62, SD = 1.19) were seen as less moral than those who told unequivocal prosocial lies (M = 4.54, SD = 0.64), $t(94) = 3.12$, $p < .01$, $d = 0.64$.

We also found a main effect of deception, $F(1, 188) = 4.94$, $p = .02$, $\eta_p^2 = .03$, such that participants generally believed that Senders were more moral when they told the truth ($M_{Honesty}$ = 4.47, $SD_{Honesty}$ = 1.16 vs. M_{Lying} = 4.06, SD_{Lying} = 1.52) There was no main effect of lie type, $p >$ .90.

Positive affect. A two-way ANOVA revealed a significant Deception x Lie Type interaction, $F(1, 188) = 12.23$, $p < .001$, $\eta_p^2 = .06$. Targets experienced more positive affect in response to unequivocal prosocial lies (M = 3.32, SD = 1.90) compared to honesty (M = 2.67, SD = 1.35), $t(92) = 1.92$, $p = .06$, $d = .40$. However, targets experienced less positive affect in response to paternalistic lies (M = 2.80, SD = 1.39) than to honesty (M = 3.71, SD = 1.50), $t(96) = 3.11$, $p < .01$, $d = .63$. There was no main effect of deception or lie type (ps > .20).

Robustness check: Perceived deception. One potential alternative account for our results is that paternalistic lies are perceived as more deceptive than unequivocal prosocial lies. To test this, we ran a two-way ANOVA with deception and lie type included as factors, using perceived deception (our manipulation check) as the dependent variable. In addition to a main effect of deception, $F(1, 188) = 321.13$, $p < .001$, $\eta_p^2 = .63$, we also found a significant Deception x Lie Type interaction, $F(1, 188) = 30.27$, $p < .001$, $\eta_p^2 = .15$. In the unequivocal prosocial lie condition, lying had a smaller effect on perceived deception (M_{Lying} = 5.03, SD_{Lying} = 1.63 vs. $M_{Honesty}$ = 2.84, $SD_{Honesty}$ = 1.12), $t(93) = 8.51$, $p < .001$, $d = 1.56$, relative to the paternalistic lie condition (M_{Lying} = 6.05, SD_{Lying} = 0.87 vs. $M_{Honesty}$ = 1.91, $SD_{Honesty}$ = 1.05), $t(97) = 16.85$, $p <$

.001, $d = 1.28$. We found no main effect of lie type ($p > .9$). These findings suggest that unequivocal prosocial lies were seen as less deceptive than paternalistic lies.

To rule out the possibility that moral judgments of unequivocal prosocial lies and paternalistic lies were driven by this difference in perceived deception, we ran a model to examine the Deception x Lie Type interaction, controlling for perceived deception. The Deception x Lie Type interaction remained significant in this model, $B = 0.78$, $p = .02$. Moral judgments were also significantly predicted by perceived deception, $B = -0.54$, $p < .001$, such that higher perceived deception was associated with lower moral judgments of the Sender. A full regression table for these analyses is available in the Supplementary Materials.

Discussion

Study 1 documents three main results. First, individuals who told paternalistic lies were seen as less moral than those who told honest statements. Second, receiving a paternalistic lie decreased targets' positive affect. Finally, paternalistic lies were judged differently than unequivocally prosocial lies; whereas unequivocal prosocial lies boosted positive affect and improved moral judgments relative to truth telling, paternalistic lies had the opposite consequences.

Study 2

Study 2 builds upon Study 1's results in three ways. First, in Study 2, we aimed to conceptually replicate Study 1's finding that those who tell paternalistic lies are viewed as less moral and elicit less positive affect than those who are honest. Here, we investigated lies with different types of outcomes. Whereas Options A and B in Study 1 were gambles with different risk profiles, Options A and B in Study 2 were gift cards for healthy or unhealthy food. This setup also mirrors the decision of whether to tell a paternalistic lie: for instance, a mother might

falsely exaggerate the negative consequences of eating candy for breakfast in order to coerce her child into making healthier choices. In Study 2, participants learned that Senders were faced with the choice of whether to tell the truth or to lie to endow the target with either of two gift cards for food, both of which he/she may like, but that differ in healthiness. Importantly, this design involves a decision in which the Sender must make assumptions about the best interests of the Receiver.

Second, in Study 2, participants received an explicit statement in the game instructions that the Sender had no stake in the game—that is, that the Sender would not receive a bonus regardless of the Receiver's choice of heads or tails. This is an important detail because it removes any lingering doubt about whether the Sender has selfish motivations for lying. Because it is clear that the Sender had no monetary incentive to lie, participants may be more apt to recognize benevolent motives for lying.

A final difference from Study 1 is that in Study 2, we used a larger sample size in order to increase statistical power and confidence in our results.

Procedure and Materials

We received 198 complete responses on Amazon Mechanical Turk (Mturk). Two hundred five participants began the experiment, but seven participants were automatically excluded from the experiment for failing the comprehension check. We also excluded nine participants who failed an attention check at the beginning of the survey, leaving a final sample of 189 participants (46.0% female; M_{age} = 33).

Participants were in the role of Receiver and were given the same instructions as they received in Study 1, with one exception. At the end of the instructions, participants were told that the Sender would earn no bonus, regardless of their choice of heads or tails. This statement was

included to minimize heterogeneity in inferences about the Sender's prosocial intensions, as well as in expectations about the Sender's payoffs, which were not specified in Study 1.

In this 2(Deception: honesty vs. paternalistic lying) x 2(Choice Set: choice set 1 vs. choice set 2) between-subjects design, participants were randomly assigned to receive an honest or dishonest message from the Sender. We followed the same procedure outlined in Study 1, except that the outcomes associated with Options A and B (i.e., the choice sets) were now either 1 lottery ticket for a $25 McDonalds gift card or 1 lottery ticket for a $25 Whole Foods gift card (see Table 3.2). A pilot study with a sample drawn from the same population ($N = 96$) revealed that participants would be equally satisfied receiving either of these gift cards, $t(190) = 1.05$, $p > .20$. We thus used these two gift cards for Study 2.

Dependent variables. After learning the veracity of the Sender's message and seeing the Sender's private information, participants answered a series of questions aimed to assess their judgments of the Sender. Participants indicated their agreement or disagreement (1 = *strongly disagree*, 7 = *strongly agree*) with nine questions about the Sender's morality ($\alpha = .96$): "I trust the Sender"; "The Sender is caring"; "The Sender is benevolent"; "The Sender is selfish" (reverse-scored); "The Sender is empathic"; "The Sender is trustworthy"; "The Sender is ethical"; "The Sender is immoral" (reverse-scored); and "the Sender is a good person."[19]

[19]In addition, we included the following exploratory items to assess mechanism (1 = *strongly disagree*, 7 = *strongly agree*): "The Sender wanted to help me"; "The Sender didn't care about what was best for me" (reverse-scored); "The Sender was making assumptions about my preferences"; "The Sender made the wrong decision for me" (reverse-scored); and "The Sender did what was right." However, mediation analysis with these items is included in the

We also measured participants' positive affect in response to the Sender. Participants received the same prompt as in Study 1, which asked them to indicate the extent to which they felt happy and grateful "in response to the Sender's behavior" (1 = *not at all*, 7 = *very much*; r = .89).

On the same page in which the dependent variables were assessed, participants saw a summary of the actions taken in the game. All subsequent experiments contained the same summary at the time the dependent variables were assessed.

Results

Moral character. Participants viewed Senders as less moral when they told a paternalistic lie (M = 3.50, SD = 1.34) than when they told the truth (M = 5.26, SD = 0.90), $t(187) = 10.62$, $p < .001$, $d = 1.55$.

Positive affect. Paternalistic deception also had a significant effect on positive affect. Participants who received a paternalistic lie (M = 2.84, SD = 1.84) reported less positive affect than those who were told the truth (M = 4.81, SD = 1.68), $t(187) = 7.69$, $p < .001$, $d = 1.19$.

Discussion

Study 2 provides further evidence for an aversion to paternalistic deception. In this study, we extended our investigation to lies that promote or inhibit specific consumptions habits and found that paternalistic lies again harmed moral judgments and decreased positive affect. A

Supplementary Materials because (a) the item "The Sender did what was right" is conceptually similar to items assessing moral judgments, and (b) the items used to assess mechanism here are different from those in Studies 3-5, where we implemented a consistent set of mediation items that were more conceptually distinct from the dependent variables measured.

paternalist may believe a person ought to choose healthy food or unhealthy food, the basis of which depends on the paternalist's own ideas about what is best for the target. Our results suggest that these types of lies would not be well-received if uncovered.

Study 3

In Study 3, we expanded our investigation in two ways. First, we investigated potential mechanisms underlying targets' moral judgments of paternalistic lies. Specifically, we measured the degree to which targets question the motivations of deceivers, the degree to which targets perceive the deceiver as violating their autonomy, and the degree to which deceivers are perceived as inaccurately predicting targets' preferences.

In addition, we examined lies with another type of tradeoff: intertemporal monetary payoffs. Many acts of paternalistic deception involve making an intertemporal choice on behalf of others. For example, when deciding whether to give overly positive feedback to a colleague on a poor performance, one faces the choice of whether to provide a short term gain (i.e., inflate the positive feedback to avoid causing emotional harm) or long term gain for the other (i.e., give honest feedback in hopes of improving their future performance).

Procedure and Materials

Five hundred fifty participants began our experiment on Mturk, but 36 participants failed the comprehension check and were automatically eliminated from the study. Zero participants failed the attention check, so we used all 534 complete responses in our analyses (46.9% female; $M_{age} = 33$).

As in Study 1, we randomly assigned participants to one of eight experimental conditions in a 2(Deception: honesty vs. lying) x 2(Lie Type: paternalistic lie vs. unequivocal prosocial lie) x 2(Choice Set: choice set 1 vs. choice set 2) between-subjects design. The description of the

Sender's information and the procedure for revealing the Sender's deception was identical to that given in Study 2. The main change we made was in the outcomes associated with Options A and B. In the paternalistic lie conditions, Options A and B resulted in the Receiver getting "1 lottery ticket for the chance to win $10 TODAY," or "1 lottery ticket for the chance to win $30 3 MONTHS FROM NOW." We ran two separate pretests on Mturk ($N = 59$, $N = 155$) using the matching method to elicit time preferences (Hardisty, Thompson, Krantz, & Weber, 2013). Both pretests revealed that $30 was the median amount participants reported would make them indifferent between receiving that amount in 3 months and $10 today. In the unequivocal prosocial lie conditions, Options A and B resulted in 1 or 2 tickets for one of these two lotteries, respectively (counterbalanced across participants).

Dependent variables. After learning the veracity of the Sender's message and the Sender's information, participants provided their moral judgments of the Sender ($\alpha = .96$). This scale was identical to that used in Study 2, except that the item "I trust the Sender" was not included, given its redundancy with the item "The Sender is trustworthy."

On the next survey page, participants evaluated their positive affect in response to the Sender's behavior using the same items we used in Study 2 (happy, grateful; $r = .79$).

Finally, participants answered questions designed to assess our proposed mechanisms: perceived benevolent intent, perceived autonomy violations, and inaccurate prediction of preferences. To measure perceived benevolent intent, participants indicated their agreement with the statement, "The Sender was trying to do what he/she thought was best for me" (reverse-scored). We assessed perceived autonomy violations directly by asking participants to rate their agreement with the following statement: "The Sender violated my autonomy." We also measured whether Receivers believed Senders inaccurately predicted their preferences with the item, "The

outcome I wanted was not the one the Sender thought I wanted." All items were displayed in a randomized order and were on a 1 to 7 scale (1 = *strongly disagree*, 7 = *strongly agree*).[20]

Results

Moral character. A two-way ANOVA revealed a significant Deception x Lie Type interaction, $F(1, 530) = 65.95$, $p < .001$, $\eta_p^2 = .11$. Participants judged Senders who told paternalistic lies ($M = 3.67$, $SD = 1.45$) as significantly less moral than those who were honest ($M = 5.40$, $SD = 0.94$), $t(265) = 11.45$, $p < .001$, $d = 1.40$. In contrast, when the benefits of lying were unequivocal, honesty did not have a significant effect on participants' moral judgments of Senders ($p > .20$). Although unequivocal prosocial lies were not perceived to be significantly more moral than truth-telling in this study, the results directionally support our hypotheses and

[20]In addition, we included several exploratory items to assess potential alternative explanations. We measured the extent to which Senders made assumptions about the preferences of the target ("The Sender was making assumptions about my preferences"); the extent to which they acted based on their own preferences ("The Sender made his/her decision based on his/her own preferences"); and whether targets perceived that the Sender attempted to exert influence over them ("The Sender was trying to influence me"). In this study as well as in Studies 4 and 5, there were no significant indirect effects of the three latter items, and inclusion of these items in mediation models did not alter results. Based on the guidance of the review team, we focus our mediation analyses on the first three items (perceived benevolent intentions, perceived autonomy violation, inaccurate prediction of preferences) in the main manuscript. We report mediation results with all items in the Supplementary Materials.

past research (e.g., Levine & Schweitzer, 2014); $M_{\text{unequivocal prosocial lie}}$ = 4.70, *SD* = 1.66 vs. M_{truth} = 4.50, *SD* = 1.30). In addition, those who told paternalistic lies (*M* = 3.67, *SD* = 1.45) were seen as less moral than those who told unequivocal prosocial lies (*M* = 4.70, *SD* = 1.66), *t*(264) = 5.36, *p* < .001, *d* = 0.66.

There was also a main effect of deception, *F*(1, 530) = 41.73, *p* < .001, η_p^2 = .07, such that participants who received a dishonest message (*M* = 4.17, *SD* = 1.64) judged Senders as less moral than those who received an honest message (*M* = 4.94, *SD* = 1.22), *t*(532) = 6.10, *p* < .001, *d* = 0.53. There was no main effect of lie type (*p* > .60).

Positive affect. A two-way ANOVA revealed a significant Deception x Lie Type interaction, *F*(1, 530) = 90.28, *p* < .001, η_p^2 = .17. Participants reported experiencing significantly less positive affect in response to paternalistic lies (*M* = 3.06, *SD* = 2.02) than honesty (*M* = 5.28, *SD* = 1.55), *t*(265) = 10.04, *p* < .001, *d* = 1.23. In contrast, participants reported experiencing significantly more positive affect in response to unequivocal prosocial lies (*M* = 4.60, *SD* = 2.17) than honesty (*M* = 3.59, *SD* = 2.05), *t*(265) = 3.92, *p* < .001, *d* = 0.48. These results are shown in Figure 3.2.

There was also a significant main effect of deception, *F*(1, 530) = 12.54, p < .001, η_p^2 = .02. Participants reported more positive affect overall when they received an honest message (*M* = 3.81, *SD* = 2.23) rather than a dishonest one (*M* = 4.42, *SD* = 2.01), *t*(532) = 3.26, *p* < .01, *d* = .28. There was no main effect of lie type (*p* > .60).

Mediation. We entered the three focal mechanism items (perceived benevolent intentions: "The Sender was trying to do what he/she thought was best for me"; perceived autonomy violation: "The Sender violated my autonomy"; inaccurate prediction of preferences: "The outcome I wanted was not the one the Sender thought I wanted") simultaneously into a

multiple-mediation model using bootstrapping with bias-corrected confidence estimates (Preacher & Hayes, 2004; 2008). We ran a moderated mediation model with 10,000 resamples using deception as the independent variable, lie type as the moderator, and moral character as the dependent variable (PROCESS Macro for SPSS, Model 7, Hayes, 2016).

Results of the mediation analyses are presented in Table 3.3. These results suggest that at least three specific processes underlie moral judgments of paternalistic lies. First, targets believed that paternalistic deceivers did not have benevolent intentions. Specifically, paternalistic lies decreased participants' beliefs that the Sender was trying to do what was best for them, $B = -1.85$, $p < .001$. Second, targets believed that the Sender violated their autonomy, $B = 1.00$, $p < .001$. Finally, targets did not believe that the Sender accurately predicted their preferences, $B = 1.33$, $p < .001$. All three of these judgments in turn were significantly associated with moral judgments of the Sender (perceived benevolent intentions: $B = 0.58$, $p < .001$; perceived autonomy violation: $B = -0.49$, $p < .001$; inaccurate prediction of preferences: $B = -0.35$, $p < .001$), and there was a significant indirect effect of paternalistic lies on moral judgments through each of the three mediators (perceived benevolent intentions: 95% CI [-1.06, -.62]; perceived autonomy violation: 95% CI [-.28, -.10]; inaccurate prediction of preferences: 95% CI [-.21, -.05]).

Importantly, we found significant evidence for moderated mediation for each of these three processes. As mentioned, a decrease in the belief that the Sender had benevolent intentions (i.e., was trying to do what was best for the target) partially mediated the decrease in perceived moral character resulting from paternalistic lies. In contrast, unequivocal prosocial lies *increased* belief in benevolent intentions of the Sender, $B = 1.32$, $p < .001$, and this belief partially mediated the positive effect of unequivocal prosocial lying on perceived moral character (95%

CI [.37, .82]). We found the same pattern for beliefs about whether the Sender correctly anticipated the outcome the target wanted: while paternalistic lies led targets to believe that Senders were not accurately predicting their preferences, unequivocal prosocial lies increased the belief that Senders *were* accurately predicting their preferences, $B = -0.99$, $p < .001$, which in turn led to more favorable moral judgments (95% CI [.03, .16]). Finally, we found that while targets viewed paternalistic lies as autonomy violations, they did not view unequivocal prosocial lies as such ($p > .25$, 95% CI [-.05, .08]).

Discussion

Study 3 provides further evidence for the results of Studies 1 and 2 and also lends support for the three mechanisms underlying the effect of paternalistic lies on moral judgments. First, beliefs that the Sender did not have benevolent intentions partially explained the effect of paternalistic lies on moral judgments. When the benefits of lying were subjective—that is, in the paternalistic lie conditions—individuals perceived that deceivers did not have targets' interests in mind. When the benefits of lying were obvious, as was the case in the unequivocal prosocial lie conditions, participants perceived that deceivers *did* have their best interests in mind. Because it was reasonable to expect that all targets would prefer two lottery tickets over one lottery ticket for the same outcome, individuals did not doubt the motives of Senders who lied to obtain this outcome for targets.

We also found evidence that perceptions of autonomy violation partially explained the effect of paternalistic lies on moral judgments. These results suggest that individuals believe that paternalistic lies send a coercive signal about the desire to control the deceived party. A related interpretation is that paternalistic lies represent a restriction of the "freedom" to have an undistorted view of the world—a view that is afforded by the truth. Interestingly, when a lie

provides individuals with clear benefits over the truth, this freedom is no longer a priority, as unequivocal prosocial lies were not seen as autonomy violations.

Moreover, we obtained evidence for a third mechanism: those who told paternalistic lies were perceived as inaccurately predicting targets' preferences. This finding is particularly striking given that we counterbalanced choice set, or the outcomes that were paired with honesty and dishonesty. Participants thought senders chose incorrectly for them when they lied, regardless of which outcome was associated with the lie. This suggests that receiving an outcome via paternalistic lying may have decreased the attractiveness of the outcome itself, consistent with reactance theory (Brehm et al., 1966; Miron & Brehm, 2006). We explored this possibility further in Study 4.

Study 4

Thus far, we have shown that paternalistic lies lead to harsher moral judgments of deceivers, that these lies decrease positive affect amongst targets, and that these effects are driven by doubts about the benevolent motivations of deceivers, the perception the paternalistic lies violate the targets' autonomy, and the perception that paternalistic deceivers inaccurately predicted targets' preferences (Studies 1-3). We also showed that these results are unique to paternalistic lies, which require the deceiver to make subjective judgments about what is beneficial for the target.

In Study 4, we explored whether individuals' preferences for outcomes change as a result of being the target of a paternalistic lie. In Study 3, we found that targets of paternalistic lies did not believe that the deceiver had correctly anticipated their preferences. As mentioned, one explanation for this result is that the experience of being lied to influenced targets' preferences. To test this notion, we examined whether targets are less satisfied with an outcome resulting

from a paternalistic lie than they are when that same outcome is obtained via honesty. Shedding light on this issue has important implications for understanding responses to paternalistic lies from policymakers. Sometimes policies are put in place via dishonest means. For instance, a government might monitor its citizens' personal data under the guise of preventing a terrorist threat, but might also plan to use that data to target other crimes. Examining outcome satisfaction allows us to make claims not only about how targets might respond to these policymakers, but also how they feel about the policies themselves.

This experiment also investigated the moderating effect of individual preferences. Although Study 3 demonstrated that participants believed Senders incorrectly predicted their preferences, it remains unclear whether this effect was driven by a shift in preferences as a result of paternalistic lies, or by the fact that many participants happened to receive their less preferred outcome when they were lied to. It is possible that targets who actually received their preferred outcome may reward rather than penalize paternalistic deception. To test this, we conducted a two-part study in which we first measured individual preferences for the outcomes that would be used in the Deception Game. Then, after a period of time had elapsed, participants played the Deception Game. This procedure allowed us to match targets' ex-ante preferences for outcomes to be used in the game with the outcome they actually obtained in the game to investigate this alternative account.

Procedure and Materials

We recruited adult participants from a city in the northeastern United States to participate in a study in exchange for a $10 show-up fee. Two hundred sixty-six participants began the study, but 11 failed the comprehension check and were automatically excluded from the experimenting, yielding 255 complete responses. Two participants failed an attention check prior

to the Deception Game, and 30 participants failed a second comprehension check after the Deception Game. Excluding these participants left us with a final sample of 223 participants (73.1% female, M_{age} = 20).

Before showing up to the laboratory, participants were required to fill out a short online questionnaire in which we measured individuals' preferences for the outcomes associated with Options A and B in the Deception Game. In particular, we asked participants whether they would prefer to receive $10 immediately or $30 3 months from now (dichotomous choice). We switched the more immediate option to "$10 immediately" from "$10 today" to strengthen the plausibility of the cover story to laboratory participants. Participants were told that the Sender with whom they were paired had previously completed the study; if participants were scheduled for the first experimental session of the day, it would seem implausible that the Sender could have already participated in a survey that could result in the participant receiving money that same day.

After arriving at the laboratory, participants were randomly assigned to one of four experimental conditions in a 2(Deception: honesty vs. paternalistic lying) x 2(Choice Set: choice set 1 vs. choice set 2) between-subjects design. The instructions for the Deception Game were the same as those used in Study 3, except that this time the component of the game in which participants chose "heads" or "tails" after viewing the Sender's message was eliminated. Instead, participants were told that their payment would be determined by the message chosen by the Sender, rather than their choice as the Receiver (as was the case in Studies 1-3). This change was implemented to ensure that participants could not arrive at an outcome by going against the Sender's message, which could introduce noise in the data (Sutter, 2009). That is, whereas arriving at an outcome by adhering to or going against the Sender's message might moderate

outcome satisfaction. By eliminating this possibility, we can ensure that outcome satisfaction can only be influenced by (a) preferences for the outcome received, and (b) the Sender's honest or dishonest message. In the Sender's information, we described the message as follows: "If you send a message that does (does not) correspond to the actual coin flip outcome, the Receiver can win a $10 bonus IMMEDIATELY ($30 bonus 3 MONTHS FROM NOW)."

Dependent variables. After the Deception Game, participants answered questions to assess their satisfaction with the outcomes by indicating their agreement (1 = *strongly disagree*, 7 = *strongly agree*) with the following statements: "I am satisfied with the outcome I received"; "I am unhappy about the outcome I received" (reverse-scored; r = .66). We also measured one item regarding satisfaction with the process: "I am satisfied with the process the Sender used to arrive at my outcome." We did not include this in our outcome satisfaction measure because it does not address outcomes per se. However, it follows a similar pattern to that of the other items, and our results do not change if we include it in our outcome satisfaction measure. Following our measures of outcome satisfaction, we also assessed mechanism with the same items used in Study 3.[21]

Results

Summary statistics. In Part 1 of the study, 28.7% of participants reported preferring $10 immediately and 71.3% preferred $30 3 months from now.

[21]We focus on mechanisms underlying moral judgments in the main text, consistent with Studies 3 and 5, and include mediation with outcome satisfaction as the dependent variable in the Supplementary Materials.

Outcome satisfaction. We conducted a two-way ANOVA entering deception and preferred outcome as factors. Preferred outcome was a binary variable indicating whether participants received their preferred outcome or not. One hundred nineteen participants received the outcome they preferred (53.3%); 104 participants did not (46.7%).

This analysis revealed a main effect of deception, $F(1, 219) = 20.18$, $p < .001$, $\eta^2_p = .08$, such that participants who received an honest message were more satisfied with the outcome they received ($M = 5.49$, $SD = 1.43$) than those who received a paternalistic lie ($M = 4.65$, $SD = 1.63$), $t(221) = 4.04$, $p < .001$, $d = 0.54$. Unsurprisingly, there was also main effect of preferred outcome, $F(1, 219) = 53.80$, $p < .001$, $\eta^2_p = .20$. Those who received their preferred outcome ($M = 5.70$, $SD = 1.31$) were more satisfied than those who did not ($M = 4.33$, $SD = 1.57$), $t(221) = 7.13$, $p < .001$, $d = 0.96$. Interestingly, however, there was no Deception x Preferred Outcome interaction ($p > .90$). Thus, the effect of honesty on outcome satisfaction did not differ depending on whether one's preferred outcome was received. These results are depicted in Figure 3.3.

In addition, we examined whether the same outcome was less satisfying when it was obtained through a paternalistic lie than when it was received through honesty. Indeed, participants who received the $30 in 3 months option were significantly less satisfied than when they had obtained that option via a paternalistic lie ($M = 4.73$, $SD = 1.59$) versus honesty ($M = 5.71$, $SD = 1.30$), $t(109) = 3.50$, $p < .001$, $d = 0.67$. Similarly, those who received $10 immediately via a paternalistic lie ($M = 4.56$, $SD = 1.69$) were significantly less satisfied than when they had received that outcome through honesty ($M = 5.28$, $SD = 1.52$), $t(110) = 2.37$, $p = .02$, $d = 0.45$.

Finally, we examined whether individuals were more satisfied when they received their less-preferred outcome via truth-telling or their more-preferred outcome via paternalistic lying.

Those who received their preferred outcome via lying (M = 5.30, SD = 1.46) were marginally more satisfied than those who received their non-preferred outcome via the truth (M = 4.75, SD = 1.53), $t(108) = 1.92$, $p = .06$, $d = 0.37$.

Discussion

Study 4 demonstrates that that paternalistic lies result in reduced satisfaction with outcomes obtained via those lies. These findings were also replicated in an additional experiment with a sample more than twice the size of the current sample (reported in the Supplementary Materials). These results suggest that findings in Studies 1-3 were indeed driven by distaste for paternalistic lies, rather than by dissatisfaction with the outcome obtained in the Deception Game. In addition, across all studies we found that participants' distaste for paternalistic lies (relative to honesty) held regardless of the actual outcomes participants received within the paternalistic lie choice sets (see detailed analyses in Supplementary Materials), thus providing further evidence that our results were not driven by initial preferences for outcomes not received. While receiving one's preferred outcome was a stronger predictor of outcome satisfaction than honesty in Study 4, even those who did receive their preferred outcome were less satisfied when it followed dishonesty rather than honesty.

These findings suggest that individuals experience reactance towards paternalistic lies and those who tell them. Study 3 indicated that targets believed that paternalistic liars inaccurately predicted their preferences. This result is consistent with reactance theory, whereby the attractiveness of an imposed option is decreased by its imposition (Brehm et al., 1966; Miron & Brehm, 2006). In Study 4, we obtained further evidence of this notion by observing that individuals were less satisfied with an outcome obtained via paternalistic lies, further implicating the role of reactance in responses to these lies.

These results also highlight the importance of both honesty and the perceived desirability of outcomes on satisfaction. While individual preferences for a policy, decision, or product may largely influence their satisfaction with these outcomes, the perceived honesty with which these outcomes are obtained likely plays a key role in the extent to which people are satisfied with these outcomes.

Study 5

In all studies reported thus far, we employed a version of the Deception Game in which no communication between the Sender and Receiver was permitted, except for the honest or dishonest message from the Sender. This design allows us to isolate the impact of dishonesty on participant responses, and also simulates targets' uncertainty about deceivers' motivation for lying. However, sometimes when an individual discovers that she has been the target of a lie, she may confront the deceiver. Given that the deceiver has acted in what she believes is the best interest of the target, the former is likely to directly express these good intentions in her defense. But how effective would this defense be at mitigating the target's unfavorable responses to the deceiver? To answer this question, we introduced a new component of the Deception Game in which the Sender could include a personalized message to the Receiver. We explored whether a message conveying the Sender's good intentions would moderate targets' responses to paternalistic lies.

In addition, all studies reported thus far have focused primarily on targets' perceptions of and reactions to paternalistic lies and those who tell them. Here, we introduced a behavioral measure of punishment to document the strength of targets' distaste for paternalistic deception.

Procedure and Materials

Five hundred forty-eight Mturk participants began our study, but 19 participants were automatically excluded the experiment because they failed the comprehension check. We also excluded one participant who failed an attention check at the beginning of the survey, yielding a final sample of 528 participants (47.2% female; M_{age} = 34).

We randomly assigned participants to one of eight experimental conditions in a 2(Deception: honesty vs. lying) x 2(Communication: communication vs. no communication) x 2(Choice Set: choice set 1 vs. choice set 2) between-subjects design. Those in the no communication conditions engaged in the Deception Game as described in Study 4 (Part 2). Those in the communication conditions received identical procedures, except with additional information about the Sender's ability to send a "personal communication" to the Receiver. These participants were told that the personal communication would be delivered along with the message about the outcome of the coin flip. Participants were told that this personal communication would not affect the bonus participants could earn. On the same screen that displayed the Sender's honest or dishonest message about the coin flip, participants in the communication condition received the Sender's personal communication. This communication read: "Just trying to get you the outcome I thought you'd want." As in the previous experiments, all participants viewed the Sender's private information (i.e., the Sender's deception or honesty, and the payoffs associated with these choices) before we collected our dependent variables.

Dependent variables. After viewing the Sender's information, which included intertemporal payoffs as the choice sets,[22] participants learned about the punishment decision.

[22]For this study, we ran another pretest (N = 54) with a different method to elicit time preferences. The results of this pretest suggested participants were roughly indifferent between

They were told that Senders would be entered into a lottery for a $10 bonus, and that they could take away any integer amount (between $0 and $10) from the Sender, though any money they took away would not be added to their own payment. Participants indicated the amount they chose to take away from the Sender, if any.

Next, we measured participants' moral judgments of the Sender (α = .94), as well as their positive affect (r = .79), using the same items from Study 3. Rather than asking participants to indicate their emotions in response to the Sender, here they were just asked to indicate the extent to which they felt happy and grateful "right now." Items to measure moral judgments and emotions were displayed on separate and counterbalanced survey pages. Finally, we assessed mechanisms using the same items as in Study 3.

Results

Manipulation check. To ensure that participants read the Sender's communication and understood that the expressed good intentions were genuine, we examined the effect of communication on the mechanism item, "The Sender was trying to do what he/she thought was best for me," collapsing across deception. A t-test revealed a significant difference, such that those who received the communication (M = 5.34, SD = 1.69) expressed greater belief in this statement than those who received no communication (M = 4.93, SD = 1.55), $t(526) = 2.87$, $p < .01$, d = 0.25.

Punishment. A two-way ANOVA with deception and communication included as factors revealed a main effect of deception, $F(1, 524) = 7.46$, $p < .01$, $\eta_p^2 = .01$. Those who were

receiving $30 in 3 months and $17.50 today. Thus, we used these options as the outcome pairings.

told a paternalistic lie (M = 1.97, SD = 3.46) punished Senders more than those who received an honest message (M = 1.23, SD = 2.73), $t(526) = 2.74$, $p < .01$, $d = 0.24$. Interestingly, there was no main effect of communication ($p > .60$) and no interaction ($p > .90$). We also looked at the difference in punishment between those who had been lied to with and without communication; the difference was not significant ($p > .70$). We depict these results in Figure 3.4.

Moral character. A two-way ANOVA with moral character as the dependent variable also revealed a main effect of deception, $F(1, 524) = 60.10$, $p < .001$, $\eta_p^2 = .01$. Senders who told paternalistic lies (M = 4.52, SD = 1.30) were judged as less moral than Senders who had been honest (M = 5.28, SD = 0.93), $t(526) = 7.69$, $p < .001$, $d = 0.67$. Unlike our results for punishment, there was also a significant main effect of communication, such that those who communicated benevolent intent (M = 4.74, 1.32) were perceived as more moral than those who did not (M = 4.31, SD = 1.25), $t(526) = 2.62$, $p < .01$, $d = 0.34$. This effect held when comparing those who received a lie with communication (M = 4.74, SD = 1.32) to those who received a lie with no communication (M = 4.31, SD = 1.25), $t(266) = 2.78$, $p < .01$, $d = 0.34$. There was no Deception x Communication interaction ($p > .10$).

Positive affect. A two-way ANOVA examining the effects of deception and communication on affect yielded results analogous to those for punishment; there was a significant main effect of deception, $F(1, 524) = 32.90$, $p < .001$, $\eta_p^2 = .06$, such that dishonesty (M = 4.09, SD = 1.73) resulted in less positive affect than honesty (M = 4.91, SD = 1.58), $t(526) = 5.74$, $p < .001$, $d = 0.50$. There was neither a main effect of communication nor a Deception x Communication interaction (ps > .5). Furthermore, there was no effect of communication on positive affect among those who had received a dishonest message ($p > .9$).

Mediation. As in Study 3, we assessed the mechanisms underlying moral judgments of paternalistic lies. We ran a moderated mediation model, with deception as the independent variable, communication as the moderator, and moral character as the dependent variable (PROCESS Macro for SPSS, Model 7, Hayes, 2016). We entered all mechanism items (perceived benevolent intentions: "The Sender was trying to do what he/she thought was best for me"; perceived autonomy violation: "The Sender violated my autonomy"; inaccurate prediction of preferences: "The outcome I wanted was not the one the Sender thought I wanted") simultaneously into a multiple-mediation model using bootstrapping with bias-corrected confidence estimates (Preacher & Hayes, 2004; 2008).

Results of the mediation analyses are presented in Table 3.4. We find significant evidence for mediation for the same three mechanisms identified in Study 3. Furthermore, we find no evidence for moderated mediation. The same mechanisms drove perceptions of moral character in both the communication and the no communication conditions.

Specifically, in both conditions, paternalistic lies resulted in decreased beliefs that the Sender had prosocial intentions (communication: $B = -0.52$, $p < .05$; no communication: $B = -0.66$, $p < .001$). Perceptions of the Sender's prosocial intentions were significantly associated with moral judgments (communication: $B = 0.52$, $p < .001$; no communication: $B = 0.54$, $p < .001$), and there was a significant indirect effect of paternalistic lies on moral judgments through this mechanism (communication: 95% CI[-.41, -.06]; no communication: 95% CI[-.45, -.12]). This result is a testament to the robustness of the skepticism about deceivers' benevolent intentions resulting from paternalistic lying. Although communicating benevolent intent improved moral judgments of Senders relative to no communication, lying increased the belief

that Senders were not acting in targets' best interests even amongst those who received the communication. This belief in turn led to lower judgments of moral character.

Furthermore, in both conditions, targets believed that the Sender violated their autonomy (communication: $B = 0.38$, $p < .05$; no communication: $B = 0.42$, $p < .05$) and thought that the Sender did not accurately predict their preferences (communication: $B = 0.98$, $p < .001$; no communication: $B = 0.71$, $p < .001$). Moral judgments were significantly predicted by both perceived autonomy violation (communication: $B = -0.44$, $p < .001$; no communication: $B = -0.38$, $p < .001$) and perceived inaccurate predictions of preferences (communication: $B = -0.26$, $p < .001$; no communication: $B = -0.31$, $p < .001$), and there was a significant indirect effect of paternalistic lies on moral judgments through each of these mechanisms for both those in the communication and no communication conditions (autonomy violation, communication: 95% CI[-.11, -.01]; no communication: 95% CI[-.12, -.01]; inaccurate prediction of preferences, communication: 95% CI[-.17, -.05]; no communication: 95% CI[-.14, -.02]).

Discussion

When deception is uncovered, a common response of the deceiver may be to defend her actions, explaining that she lied because she believed it was in the target's best interest. In Study 5, we tested the effectiveness of this type of defense. While communication of benign intentions did improve judgments of Senders' moral character, it had no effect on punishment of deceivers or on targets' emotional responses to deception. Moreover, targets believed deceivers were not prosocially motivated, viewed lying as an autonomy violation, and thought deceivers inaccurately predicted their preferences, even when the deceiver tried to communicate good intentions.

Study 6

In Studies 1-5, we provided consistent evidence of an aversion to paternalistic lies. However, one potential criticism of these studies is that the Deception Game, while well-controlled, does not fully capture the essence of paternalistic lies. Though the subjective nature of the benefits of paternalistic lies in the game are analogous to the uncertainty associated with real-world outcomes of paternalistic lies, the abstract framing of the game is quite dissimilar to the real-world contexts in which paternalistic lies are told. Furthermore, interactions in the game were between strangers, whereas paternalistic lies in everyday life often occur between friends, colleagues, romantic partners, and other relationships in which both parties are at least acquainted with one another.

Considering these issues, in Study 6, we implemented a different methodology that allowed us to measure judgments of paternalistic lies in a more externally valid setting. Here, participants read several vignettes in which they were asked to imagine that they discovered they had been lied to. In each vignette, we manipulated (a) whether the interests of the target were known or unknown to the deceiver, and (b) whether the deceiver communicated his/her benevolent intent to the target. We included both paternalistic and unequivocal prosocial lies in this study to determine whether individuals indeed respond to these lies differently in more externally valid contexts. Specifically, we sought to provide further evidence of Study 1 and 3's findings that those who tell paternalistic lies (i.e., when the interests of the target are unknown) are viewed as less morally acceptable than those who tell unequivocal prosocial lies (i.e., when the interests of the target are known). Additionally, we extended Study 5's investigation of the effects of communication on judgments of paternalistic deceivers in order to determine whether communicating benign intent has differential effects when there is an existing relationship between the target and deceiver.

Procedure and Materials

We received 395 complete responses from Mturk. Three participants failed an attention check at the beginning of the study and were thus excluded. We also excluded three responses from participants who had already taken the survey (though the original responses of these participants were retained). This left a final sample of 388 (42.8% female, $M_{age} = 38$).

Participants were randomly assigned to one of four conditions in a 2(Lie Type: paternalistic lie vs. unequivocal prosocial lie) x 2(Communication: communication vs. no communication) between-subjects design. Within each condition, all participants read three vignettes that were displayed in a randomized order, with page breaks separating each vignette. In these vignettes, participants were asked to imagine that they had been the target of a paternalistic or unequivocal prosocial lie (depending on condition). Whereas in Studies 1-5 we manipulated paternalistic deception by altering whether the benefits of lying were subjective or objective, in Study 6 we directly manipulated the degree to which the deceiver was aware of the target's preferences, while holding constant the amount of time the deceiver and target knew each other in each vignette. We also manipulated whether the deceiver did or did not communicate his/her intentions to help the target by lying (depending on condition). For example, one scenario read as follows:

> *You and your friend Jill are out to dinner at Jill's favorite restaurant. You are trying to lose weight and eat healthy. You ask what Jill recommends. She says that the signature salad is her favorite item on the menu. Weeks later you learn that Jill lied and that her favorite menu item is actually the double cheeseburger.*

Paternalistic lie: *You and Jill have been friends for about 6 months. You two had never discussed whether you desired to lose weight and avoid temptation or to indulge in tasty but unhealthy foods.*

Unequivocal prosocial lie: *You and Jill have been friends for about 6 months. You two had discussed you desire to lose weight and avoid temptation or to indulge in tasty but unhealthy foods.*

Communication: *Jill tells you that she lied because she wanted you to eat healthy.*

No communication: [*No additional information*].

The other two vignettes, which depict lies from a coworker and a doctor, are reprinted in the Appendix.

After each vignette, participants provided moral judgments of the deceiver, as well rated the positive affect they expected to experience in response to the deceiver's behavior. The items and scales used to measure moral judgments were the same as those in Studies 3 and 5. For moral judgments, the prompt read, "Please indicate the extent to which the following words characterize [Jill] from the scenario above. [Jill] is…" For affect, the prompt read, "If you were actually the person in the above scenario, please indicate the extent to which you would experience the following emotions in response to [Jill's] behavior." Each vignette was displayed to participants as they made their ratings.

Results

In Study 6, we were interested in examining the effect of lie type, communication, and their interaction on judgments of moral character and positive affect. We therefore report the

results of 2 (Lie Type: unequivocal prosocial lie vs. paternalistic lie) x 2 (Communication: communication vs. no communication) ANOVAs on judgments of moral character and affect collapsed across vignettes. Mixed model ANOVAs that include the effects of vignette are included in the Supplementary Materials. However, inclusion of vignette in the models does not moderate our results.

Moral character. There was a significant effect of lie type on judgments of moral character, $F(1, 384) = 67.56$, $p < .001$, $\eta_p^2 = .15$. Those who imagined they were targets of unequivocal prosocial lies ($M = 4.54$, $SD = 0.66$) judged deceivers as more moral than those who were targets of paternalistic lies ($M = 3.99$, $SD = 0.66$), $t(386) = 8.20$, $p < .001$, $d = .83$. There was no main effect of communication ($p = .13$) and no Lie Type x Communication interaction ($p = .18$). We also tested whether communication had an effect within each lie type; there was a marginally significant effect of communication for those in the paternalistic lie conditions, $t(192) = 2.02$, $p = .05$, $d = .29$. Communication marginally improved moral judgments of those who told paternalistic lies ($M_{communication} = 4.08$, $SD_{communication} = 0.67$ vs. $M_{no\ communication} = 3.89$, $SD_{no\ communication} = 0.64$). There was no effect of communication for those in the unequivocal prosocial lie conditions ($p > .25$). These results are displayed in Figure 3.5.

Positive affect. Similar results were obtained for positive affect. There was a significant effect of lie type, $F(1, 384) = 73.62$, $p < .001$, $\eta_p^2 = .16$, such that those who imagined they were targets of unequivocal prosocial lies ($M = 3.81$, $SD = 1.12$) reported more positive affect than those who were targets of paternalistic lies ($M = 2.83$, $SD = 1.10$), $t(386) = 8.57$, $p < .001$, $d = .87$. There was no effect of communication ($p > .25$), and no interaction ($p > .25$). We also examined the effect of communication within each lie type; the effect of communication was not

significant for either those who received a paternalistic lie or those who received an unequivocal prosocial lie ($ps > .10$).

Discussion

Study 6 provides evidence for the external validity of individuals' aversion to paternalistic lies. Using a design that depicted realistic contexts and relationships in which paternalistic lies are told, we replicated Studies 1 and 3's findings that paternalistic lies result in harsher moral judgments than unequivocal prosocial lies.

In addition, this study offers evidence of the limitations of communication on mitigating the negative effects of paternalistic lies. In both Studies 5 and 6, communication had no effect on affect resulting from being the target of a paternalistic lie. Unlike in Study 5, however, where communication improved moral judgments of paternalistic liars, in Study 6, communication had no effects on moral judgments of those who tell paternalistic lies. Given the use of a more realistic context for studying paternalistic lies in Study 6, these results suggest that the ability of the communication of benevolent intent to reduce the harmful effects of moral judgments of paternalistic liars is limited.

Study 7

In Studies 3 and 5, we assessed the mechanism behind paternalistic lies' effects on moral judgments using mediation analysis. While consistent mediation results across these studies provides evidence for the underlying process, this type of analysis is limited by its correlational nature (Spencer, Zanna, & Fong, 2005). In Study 7, we provide stronger causal evidence for one of the mechanisms uncovered in Studies 3 and 5—perceived benevolent intent—by directly manipulating this construct.

In this experiment, we employed a vignette design similar to that of Study 6, where participants read three different vignettes depicting the telling of paternalistic lies. Here, we varied between -subjects whether the benevolent intent of the deceiver was ambiguous, or made clear to participants through a statement that provided an omniscient third-person perspective into the inner state of the deceiver. Although we find that personally communicating one's good intent was not seen as credibly signaling benevolence in the context of paternalistic lies, this study directly examines the role of perceived benevolent intent by manipulating it directly. We predicted that if targets knew for certain that deceivers lied with good intentions, this would improve moral judgments relative to when the motivations of the deceiver are more ambiguous. In Study 7, we also included new vignettes that depicted paternalistic lies told by individuals in leadership positions, thus allowing us make inferences about the effects of paternalistic lies in relatively higher-stakes contexts.

Procedure and Materials

We received 214 complete responses from Mturk. Eight participants were excluded from analyses for failing an attention check. This left a final sample of 206 (53.9% female, $M_{age} = 38$).

Participants were randomly assigned to one of two conditions: ambiguous motivation or benevolent motivation of deceivers. Within each condition, participants read three vignettes in a randomized order. As in Study 6, for each vignette participants were asked to imagine that they were the target of a paternalistic lie. In Study 6, we manipulated whether the deceiver verbally communicated benevolent intent to the target. In Study 7, we manipulated whether the deceivers' benevolent intent was clear to participants by including a statement from third-person omniscient perspective that described the deceiver's private thoughts and motivation. For example, one vignette depicted a government official who lied to constituents:

Imagine that you are at a community board meeting listening to a local government official speak. There have been rumors about a possible security threat in your city, and the government official is addressing those concerns.

The official insists that the rumors are unsubstantiated, and that there is no security threat.

Weeks later, however, news emerges that there was in fact substantial evidence of a security threat, and the government official knew about this evidence at the time of the community board meeting.

This government official had been in his/her position for around 6 months, and was unaware of your preferences and other constituents' preferences to be fully informed in the event of a threat, or to be uninformed in order to not worry.

Benevolent Motivation: *In actuality, the government official lied about the security threat because s/he believed there was nothing the public could do about the threat and that everyone would be better off not worrying. S/he was sincerely trying to do what s/he thought was best for you and the public.*

Ambiguous Motivation: [*No additional information*].

The second vignette depicted a lie from a doctor (adapted from Study 6), and the third depicted a lie from a financial advisor. These vignettes are reprinted in the Appendix.

After each vignette, participants provided moral judgments of the deceiver, using the same items and prompt as in Study 6. Each vignette was displayed to participants as they made their ratings. In addition, we included items to measure each of the three mechanisms identified in Studies 3 and 5: perceived benevolent intent, autonomy violation, and inaccurate prediction of

preferences. These items were the same as those used in Studies 3 and 5 to measure these constructs, except "the Sender" was replaced with the deceiver depicted in the vignette (i.e., the doctor, the financial advisor, the government official). The item "The [deceiver] was trying to do what s/he thought was best for me" served as a manipulation check of perceived benevolent intent. The items "The outcome I wanted was not the outcome the [deceiver] thought I wanted" and "The [deceiver] violated my autonomy" served as tests of discriminant validity—that is, if our experimental treatment indeed manipulated benevolent intent only, the manipulation should not produce changes in these items measuring other constructs.

Results

In Study 7, we sought to test the impact of benevolent intentions on moral judgments of paternalistic lies. Because there were only two between-subjects treatments in this experiment, we report results of t-tests to compare moral judgments of deceivers across conditions, collapsing across vignettes. Mixed model ANOVAs that include the effects of vignette are reported in the Supplementary Materials.

Manipulation check. A t-test indicated that our benevolent intent manipulation worked as planned. Those in the benevolent motivation condition exhibited higher scores ($M = 4.68$, $SD = 1.19$) on the item, "The [deceiver] was trying to do what s/he thought was best for me," than those in the ambiguous motivation condition ($M = 4.18$, $SD = 1.07$), $t(204) = 3.15$, $p < .01$, $d = .44$.

Moral character. There was a significant effect of the motivation manipulation on moral judgment of deceivers. Those in the benevolent motivation condition ($M = 3.76$, $SD = 0.69$) rated

deceivers as more moral than those in the ambiguous motivation condition (M = 3.53, SD = 0.65), $t(204) = 2.48$, $p = .01$, $d = .35$.[23]

Discriminant validity. As mentioned, our manipulation successfully produced increased in perceived benevolent intent. In order to assess the discriminant validity of this manipulation, we examined whether this manipulation also affected perceived autonomy violation, or the perception that the deceiver inaccurately predicted one's preferences. There were no differences across conditions for the item "The [deceiver] violated my autonomy" ($p = .14$), nor for the item "The outcome I wanted was not the one the [deceiver] thought I wanted" ($p > .25$).

Discussion

In Study 7, we provide causal evidence that perceived benevolent motivation is a mechanism underlying the effects of paternalistic lies on moral judgments. An experimental manipulation that made explicit deceivers' internal desire to benefit the target via lying improved moral judgments, relative to when deceivers' motivations were not specified. Moreover, the manipulation of benevolent intent did not influence perceived autonomy violation or perceived inaccurate prediction of preferences, thereby highlighting the discriminant validity of this manipulation and providing evidence that these three mechanisms are indeed unique constructs.

[23]As described in the Supplementary Materials, there was also a significant Motivation x Vignette interaction, $F(2, 408) = 4.60$, $p = .01$, $\eta_p^2 = .02$. The effect of motivation was significant for the government ($M_{benevolent}$ = 3.65, $SD_{benevolent}$ = 0.83 vs. $M_{ambiguous}$ = 3.27, $SD_{ambiguous}$ = 0.94; $F(1, 204) = 9.34$, $p < .01$, $\eta_p^2 = .04$) and finance ($M_{benevolent}$ = 3.64, $SD_{benevolent}$ = 0.92 vs. $M_{ambiguous}$ = 3.29, $SD_{ambiguous}$ = 0.88; $F(1, 204) = 8.19$, $p < .01$, $\eta_p^2 = .04$) vignettes, but not for the healthcare vignette ($p > .25$).

These results bolster evidence from mediation analyses in Studies 3 and 5, highlighting the importance of perceived motivation in determining responses to paternalistic lies. In addition, this study expands the contexts in which we investigate paternalistic lies. Compared to the vignettes in Study 6, those in Study 7 depict lies from individuals in leadership positions in relatively higher-sakes situation, thereby offering further evidence of the potentially detrimental effects of paternalistic lies.

General Discussion

This work adds to our understanding of deception, highlighting how responses to lies hinge on the perceived benefits afforded by lying, as well as the perceived motives of deceivers. Although targets may reward lies that yield unequivocal benefits, they penalize lies that involve others making subjective judgments about their best interests. We identify a robust distaste towards paternalistic deception across moral judgments, affect, punishment, and satisfaction with outcomes associated with lying.

Our research makes several contributions to theory on deception. First, we broaden the taxonomy of lies by introducing the construct of paternalistic lies. Although paternalistic lies are ubiquitous and have important consequences for both targets and deceivers, no prior research has examined these lies. We distinguish paternalistic lies from unequivocal prosocial lies, another class of lies that are intended to benefit others that have been studied in past work, and demonstrate how responses to paternalistic lies differ from responses to unequivocal prosocial lies.

This research also extends the growing body of research on prosocial lying. Our results identify a boundary condition of the positive effects of prosocial lying (Levine & Schweitzer, 2014, 2015), showing that paternalistic lies and unequivocal prosocial lies can yield divergent

moral judgments and affective responses. In Levine & Schweitzer's (2014, 2015) work, unequivocal prosocial lies were perceived to be benevolent. Similarly, in the present research, unequivocal prosocial lies elicited the judgment that the deceiver was truly trying to do what they thought was best for the target (see mediation results in Studies 3 and 5), which is also indicative of perceived benevolent intent. This credible signal of benevolence lead to positive judgments of moral character. In contrast, for paternalistic lies, the signal of benevolence is less credible. We find that targets do not believe that deceivers who tell a paternalistic lie were truly trying to do what they thought was best for the target, and that this diminished belief in deceivers' benevolent intent in turn drove the decrease in perceived moral character. Thus, this research highlights the theoretical and practical importance of perceived benevolence in shaping moral judgments.

In addition to identifying perceived benevolent intent as a mechanism behind negative responses to paternalistic lies, we also uncover two additional mechanisms underlying these responses: the perception that paternalistic lies violate targets' autonomy, and the perception that paternalistic liars inaccurately predicted targets' preferences. Not only do these findings shed further light on the processes that drive responses to paternalistic lies, but they also suggest that paternalistic lies can elicit reactance amongst targets. According to Miron and Brehm (2006), behavioral indicators of reactance include derogation of the agent restricting one's freedom, as well as a decrease in attractiveness of the imposed option or an increase in the attractiveness of the restricted option. In our experiments, we see evidence for both of these phenomena. Participants derogated deceivers via moral judgments (Studies 1-3, 5-7), and punishment (Study 5). Furthermore, perceived inaccurate prediction of preferences drove decreases in moral judgments (Studies 3 and 5), and paternalistic lies actually decreased satisfaction with outcomes

that were received as a result of these lies (Study 4). We also find that paternalistic lies harm affective responses—another sign of reactance (Miron & Brehm, 2006). Taken together, these findings provide the first evidence that we know of that deception can produce reactance.

Our results also present a novel application to theory on procedural justice. A widespread finding in the justice literature is the significant interaction between procedural fairness and outcome desirability, such that the relationship between procedural fairness and individuals' reactions is stronger when outcome desirability is low (Brockner & Wiesenfeld, 1996). This finding has not yet been applied to judgments of deception, yet our results fit this pattern nicely: when individuals are the target of an unequivocal prosocial lie (i.e., a lie with objectively desirable outcomes), they respond favorably, despite the arguably unfair or immoral action that was taken to produce those outcomes. When they are the target of a paternalistic lie, however, (i.e., a lie with outcomes that are not objectively desirable), they become more sensitive to the fact that they were lied to, and thus, respond harshly. Similarly to how perceptions of outcomes and procedures interact to produce individuals' reactions in an organizational context, the degree to which individuals react negatively or favorably to lies depends on the relative desirability of the outcomes associated with those lies.

Apart from its theoretical contributions, this work also has practical implications for interpersonal interactions, management, and policy-making. Leaders and policy-makers often withhold or distort the truth in the perceived best interests of their stakeholders. Although targets may respond positively when the lie is clearly favorable to them, individuals often lack full insight into others' preferences (e.g., Hsee & Weber, 1997), and there is often uncertainty about the ultimate consequences deception. Our results indicate that well-intended lies may backfire if deceivers lack sufficient insight into what is actually in targets' best interests. Targets are likely

to penalize paternalistic lies, as well as the policies, people, and products associated with them.

Relatedly, our work indicates that paternalistic lies have detrimental effects not only interpersonal perceptions, but also perceptions of outcomes resulting from these lies. Sometimes individuals need to make decisions on behalf of stakeholders that require a choice between two alternatives that have different assets and tradeoffs. For example, a government organization may be faced with the decision of whether to protect citizens' privacy, or obtain personal data to screen for a terrorist threat (e.g., Nakashima, 2016). The decision-maker may act in what she truly believes is the stakeholder's best interest, and the stakeholders' preferences for each of these options are clearly important in determining their satisfaction with the decision. However, our work suggests that the stakeholders may respond more favorably to the outcome that is delivered with transparency than to the outcome that is delivered via deception.

Our results open up several potential avenues for further research. One important area of future study would be to investigate moderators of responses to paternalistic lies to determine how opposition to these lies might be reduced. We obtained mixed evidence that communicating benign intent can soften the blow of paternalistic lies: communication did improve moral judgments of paternalistic liars in Study 5, but not in Study 6. Communication also did not decrease punishment of paternalistic liars (Study 5). However, in Study 7, knowledge of deceivers' good intentions via insight into their internal thoughts did improve moral judgments. This suggests that communication in Studies 5 and 6 may not have effectively convinced participants of deceivers' benign intentions. It may be that if communication does successfully convey deceivers' good intent, it would allay the negative effects of paternalistic lies. Given the limited effectiveness of communication in our work, future research should examine other ways

in which liars can successfully convey their benevolent intentions in order to mitigate the harmful effect of paternalistic lies.

Conversely, there are likely other factors that can exacerbate negative responses to paternalistic lies that are worthy of further investigation. In our research, we purposely structured the Deception Game such that the deceiver had no stake in the game so that we could cleanly study paternalistic lies (relative to the truth and unequivocal prosocial lies), without confounding paternalism with self-interest. Likewise, in the vignettes used in Studies 6 and 7, no ulterior motives of deceivers are mentioned. In the real world, however, deceivers may have mixed motives. For example, one tasked with delivering feedback about a poor performance may upwardly inflate this feedback to prevent causing emotional harm, but also to avoid the discomfort of an awkward situation. In this work, we find that perceived intentions of deceivers play a key role in the divergent effects of paternalistic lies and unequivocal prosocial lies. We would expect, then, that if the liar was known or believed to have ulterior motives, paternalistic lies would be penalized to an even greater extent. More research would serve well to explore this notion.

It will also be important for future work to examine the situations in which lies are more likely perceived to be paternalistic, versus unequivocally prosocial. In certain circumstances, there may be broad consensus that lying serves a target's best interests. In these cases, lies are likely to elicit positive reactions. For example, most people may agree that telling a bride she is beautiful on her wedding day is in the bride's best interest, regardless of the truth. Thus, an individual who tells such a lie may be rewarded. However, in other circumstances, there may be little consensus on whether lying is beneficial. In these circumstances, the lie will likely be perceived as paternalistic, and elicit negative reactions. For example, there may be considerable

disagreement about whether falsely telling a woman she looks beautiful on an ordinary day is in the woman's best interest. Thus, an individual who tells such a lie may be penalized. Recent research suggests that there are systematic circumstances in which lies are generally perceived to benefit targets (Levine, 2017). It will be interesting for future research to examine if judgments of paternalism are reduced in these contexts.

Another possibility for future work would be to investigate how the relationship between the deceiver and the target influences perceptions of paternalistic lies. In close interpersonal relationships, targets may trust communicators to accurately predict their preferences and may be less skeptical of their motives. In these circumstances, individuals may experience less hostility towards paternalistic lies. Consistent with this proposition, recent research suggests that perceptions of paternalistic policies hinge on trust in the policy-maker (Tannenbaum & Ditto, 2016; Tannenbaum, Fox, & Rogers, 2016). While we investigate paternalistic lies between strangers in Studies 1-5 and a variety of closer relationships in Studies 6 and 7, more research is necessary to isolate how paternalistic lies are viewed in close versus distant relationships, and how other specific features of a deceiver-target relationship may moderate responses to these lies.

A final potential avenue for future research would be to explore how the method of deception influences perceptions of paternalistic lies and those who tell them. In our research, we explore paternalistic lies in the form of a false statement from deceivers. However, there are other forms of deception that can be considered paternalistic. For example, when faced with the opportunity to tell a paternalistic lie, one can omit information in order to deceive someone for their purported benefit (i.e., lies of omission). One can also choose to change the subject of conversation, or actively choose to not disclose any information (e.g., pleading the Fifth

Amendment). Recent work suggests that opting to not disclose negative information can result in worse judgments than honest disclosure (John, Barasz, & Norton, 2016). It would be interesting for future work to test how paternalistic lies fare against these alternate modes of communication in terms of influencing social judgments of the communicator.

People are frequently faced with opportunities to engage in paternalistic deception. Though individuals might be tempted to lie with the intent to help others, the uncertainty laden in how the lie will affect the targets should give the potential deceivers pause about the decision. When the consequences of dishonesty are not unequivocally preferable to those of honesty, these parties may be better off telling the truth.

Chapter 3, in full, is a reprint of material as it appears in Organizational Behavior and Human Decision Processes, which was co-authored by Emma E. Levine and Adam Eric Greenberg in 2018. The dissertation author was the primary investigator and author of this paper.

Figures

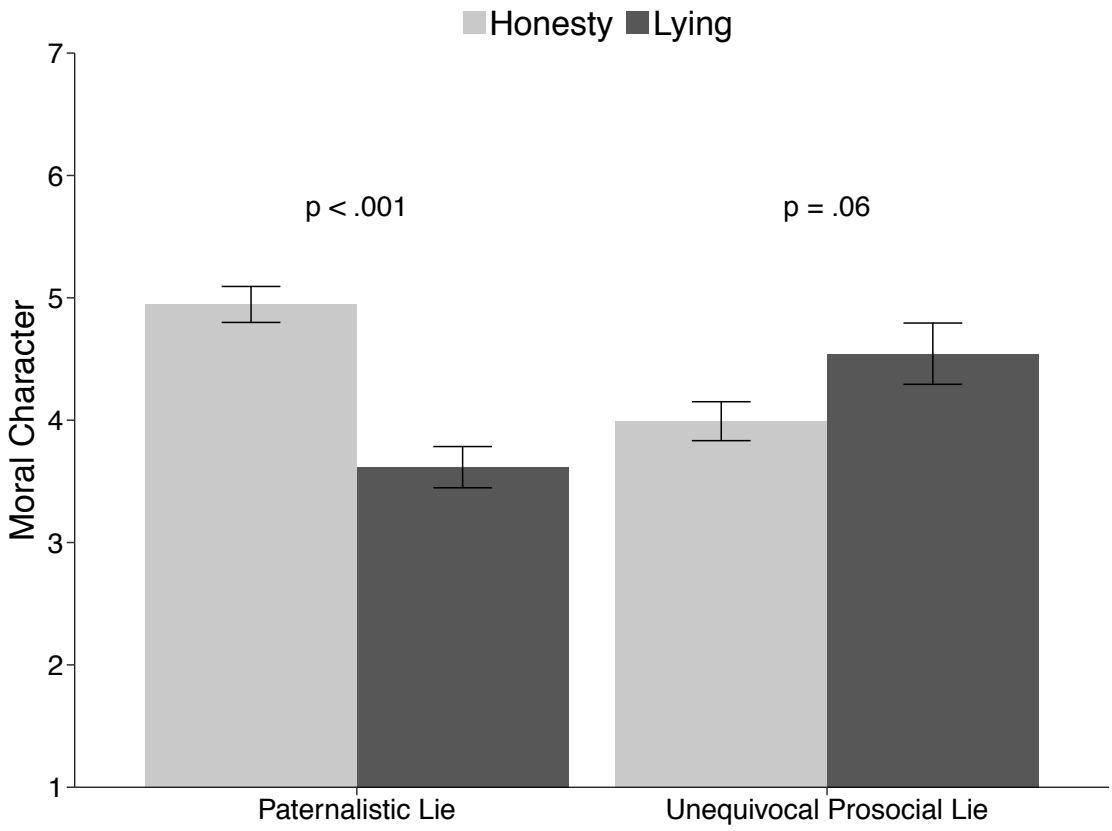


Figure 3.1: The effects of unequivocal prosocial lies and paternalistic lies on perceived moral character in Study 1. In the unequivocal prosocial lie conditions, lying and honesty were associated with 2 vs. 1 lottery tickets to the same gamble, respectively. In the paternalistic lie conditions, lying and honesty were each associated with 1 lottery ticket to either a high-risk/high-reward gamble or a low-risk/low-reward gamble. Error bars reflect +/- 1 SE.

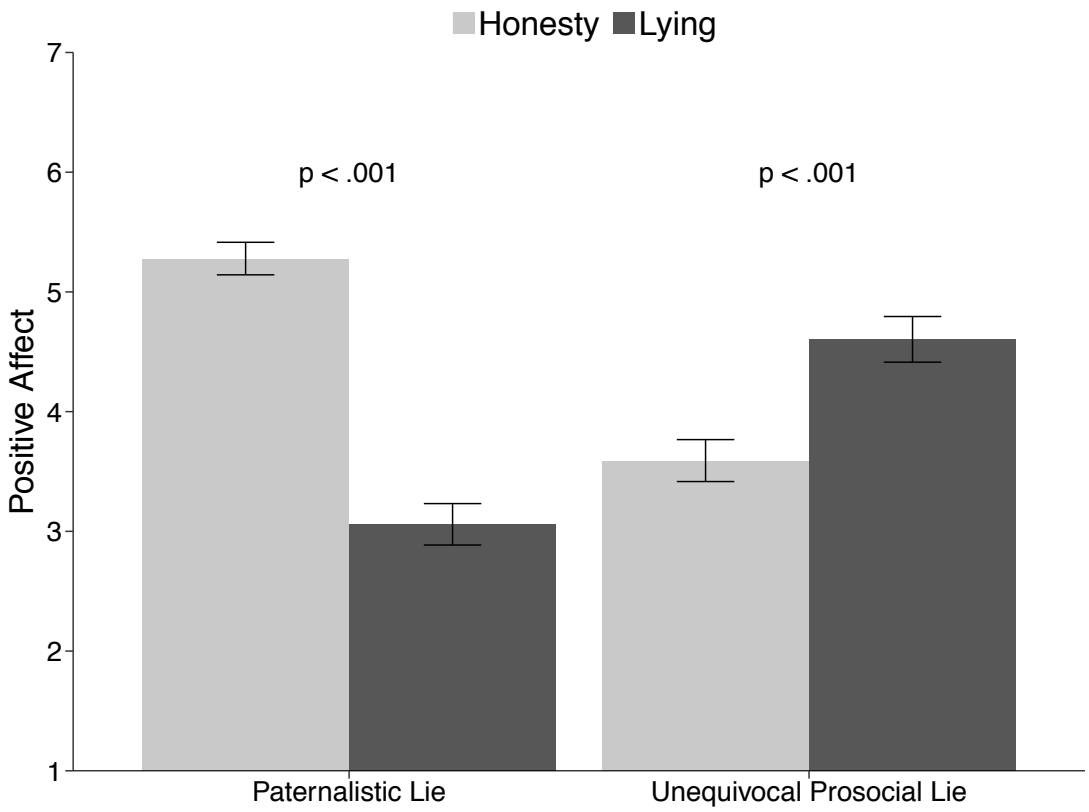


Figure 3.2: The effects of unequivocal prosocial lies and paternalistic lies on positive affect in Study 3. In the unequivocal prosocial lie conditions, lying and honesty were associated with 2 vs. 1 lottery tickets, respectively, for the same monetary outcome at the same point in time. In the paternalistic lie conditions, lying and honesty were each associated with either less money today or more money in the future (i.e., different intertemporal choices). Error bars reflect +/- 1 SE.

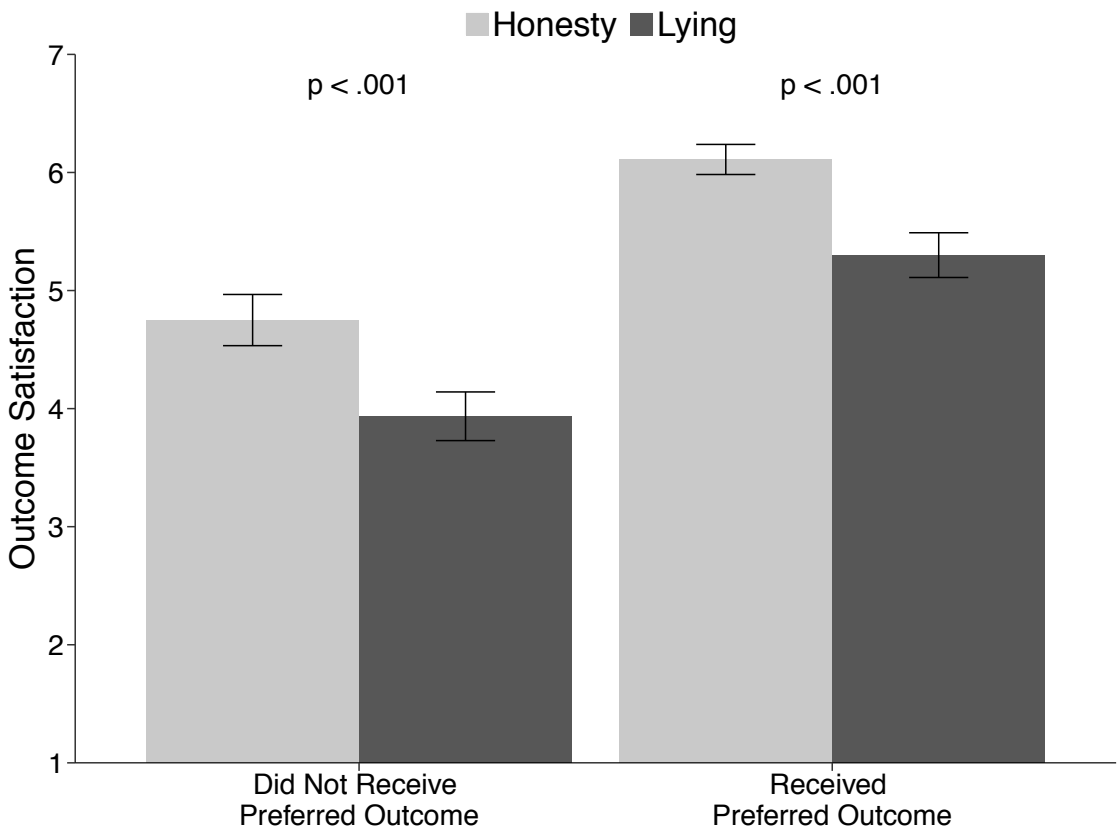


Figure 3.3: The effects of receiving one's preferred outcome and paternalistic lies on outcome satisfaction in Study 4. Receiving one's preferred outcome was dummy coded based on participants' reported preference for either "$10 immediately" or "$30 3 months from now" in Part 1 of the Study. Error bars reflect +/- 1 SE.

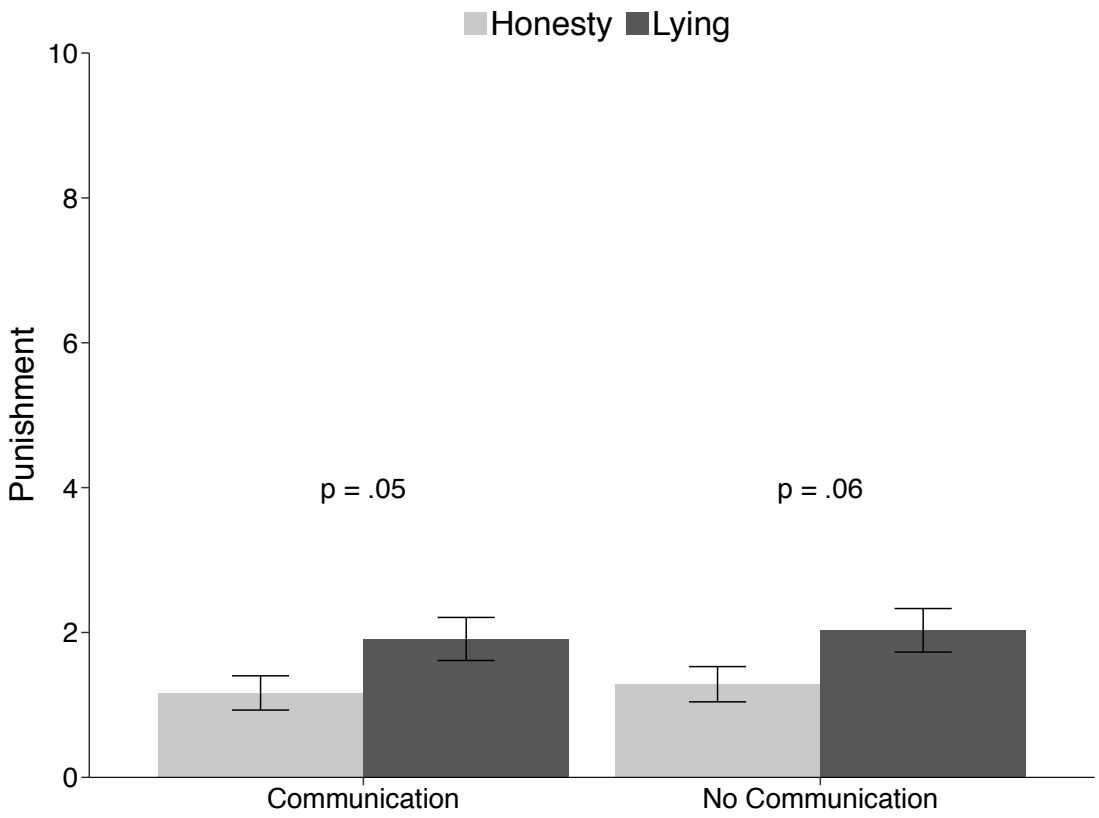


Figure 3.4: The effects of communication and paternalistic lies on punishment in Study 5. Participants in the communication condition received a personal communication from the Sender signaling benevolent intent. Those in the no communication condition received no additional communication. Error bars reflect +/- 1 SE.

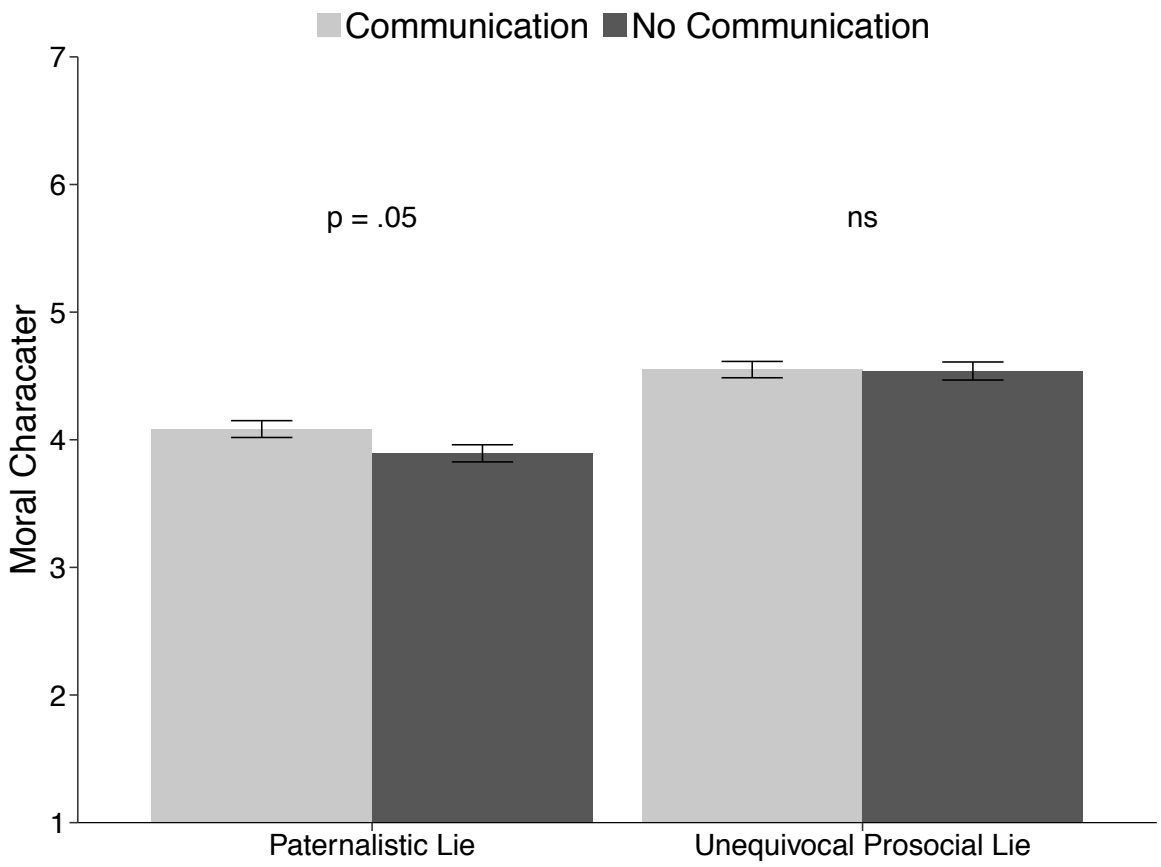


Figure 3.5: The effects of communication on perceived moral character for those who received a paternalistic lie or an unequivocal prosocial lie in Study 6. Participants in the communication viewed a statement from the deceivers depicted in the vignettes that signaled their benevolent intent. Those in the no communication condition saw no additional communication. Error bars reflect +/- 1 SE.

Tables

Table 3.1: Definitions of terms, with examples

Prosocial Lies

False statements made with the intensions of misleading and benefitting a target (Levine & Schweitzer, 2014, 2015)

Unequivocal Prosocial Lies	**Paternalistic Lies**
False statements made with the intention of misleading a target, *and are known to both the deceiver and the target to be in the target's best interests*.	False statements made with the intention of misleading and benefitting a target, *and require the deceiver to make assumptions about the target's best interests.*
Example:	**Example:**
Your spouse has terminal cancer. **You and your spouse told your doctor in the past that you both would like to remain hopeful about the prognosis rather than receive complete candor**. Your doctor falsely tells you that your spouse may be eligible for a new experimental treatment soon.	Your spouse has terminal cancer. **You and your spouse had never discussed with your doctor whether you both would like to remain hopeful about the prognosis or receive complete candor.** Your doctor falsely tells you that your spouse may be eligible for a new experimental treatment soon.

Table 3.2: Summary of the Deception Game across Studies 1-5. In each study, the payoffs associated with Options A and B were counterbalanced between-subjects.

Study	**Type of Payoffs**	**Outcomes Associated with Options A and B**	**Real World Example**
1	Gambles	50% chance of $1, 50% chance of $0 / 25% chance of $2.25, 75% chance of $0	Lying to a patient to ensure s/he chooses a low-risk medical procedure
2	Gift Cards	$25 McDonald's Gift Card / $25 Whole Foods Gift Card	Lying to a friend to ensure s/he chooses a healthy snack
3, 4, 5	Intertemporal Choice	$10, Today / $30, 3 Months from Now (Studies 3,4) $17.50, Today / $30, 3 Months from Now (Studies 5)	Lying to a client to ensure s/he saves money for the future

Table 3.3: Results of mediation analyses from Study 3. Each set of numbers signifies the lower-level and upper-level 95% confidence intervals around the indirect effect for the corresponding item in the first column. The model that was tested included all items in the first column as simultaneous mediators, deception as the IV, moral character as the DV, and lie type as the moderator. We used Hayes' (2016) PROCESS Macro for SPSS, Model 7. Bold numbers indicate confidence intervals that do not contain zero.

	Paternalistic Lies	**Unequivocal Prosocial Lies**	**Index of moderated mediation**
1. The Sender was trying to do what he/she thought was best for me	**-1.06, -.62**	**.37, .82**	**-1.77, -1.09**
2. The Sender violated my autonomy	**-.28, -.10**	-.05, .08	**-.33, -.09**
3. The outcome I wanted was not the one the Sender thought I wanted	**-.21, -.05**	**.03, .16**	**-.35, -.08**

Table 3.4: Results of mediation analyses from Study 5. Each set of numbers signifies the lower-level and upper-level 95% confidence intervals around the indirect effect for the corresponding item in the first column. The model that was tested included all items in the first column as simultaneous mediators, deception as the IV, moral character as the DV, and communication as the moderator. We used Hayes' (2016) PROCESS Macro for SPSS, Model 7. Bold numbers indicate confidence intervals that do not contain zero.

	Communication	**No Communication**	**Index of moderated mediation**
1. The Sender was trying to do what he/she thought was best for me	**-.41, -.06**	**-.45, -.12**	-.28, .18
2.. The Sender violated my autonomy	**-.11, -.01**	**-.12, -.01**	-.07, .05
3. The outcome I wanted was not the one the Sender thought I wanted	**-.17, -.05**	**-.14, -.02**	-.04, .10

Appendix

Vignettes in Study 6

A. Healthcare vignette

Imagine that your spouse has a fatal cancer. You and your spouse met with the doctor, who informed you that your spouse's existing treatment has not been effective and that the cancer has spread to your spouse's bones and brain. You know your spouse may pass away soon and you have already prepared for the worst. However, the doctor says that there is always hope and that your spouse may qualify for a new experimental treatment soon. A few weeks later, your spouse passes away. You subsequently find out that the doctor knew that your spouse was too sick to receive any experimental treatments in the future.

Unequivocal prosocial lie: *The doctor had known you and your spouse for around 6 months. He had discussed your and your spouse's preferences for negative information. He knew that you both wanted to remain hopeful and optimistic rather than receive complete candor in such dire circumstances.*

Paternalistic lie: *The doctor had known you and your spouse for around 6 months. He had never discussed your and your spouse's preferences for negative information. He did not know whether you and your spouse wanted to remain hopeful and optimistic, or whether you and your spouse wanted complete candor in such dire circumstances.*

No communication: [*no additional information*].

Communication: *The doctor tells you that he lied about the experimental treatment options because he wanted to preserve your and your spouse's hope.*

B. Feedback vignette

Imagine that you are an employee of a large consumer packaged-goods company. You have been chosen to deliver a speech to thousands of your fellow coworkers at this year's annual sales meeting. The day before the meeting, you practice your speech in front of your coworker, Nick. Nick tells you that the speech is wonderful. At the sales meeting, your speech went fine. However, several weeks later, you find out that Nick actually did not think the speech was particularly interesting or engaging when he first heard it.

Unequivocal prosocial lie: *Nick has been your coworker for about 6 months. You had told him in the past that you benefit from encouragement and reassurance rather than criticism before giving speeches.*

Paternalistic lie: *Nick has been your coworker for about 6 months. You had not discussed in the past whether you would benefit from encouragement and reassurance or criticism before giving speeches.*

No Communication: [*no additional information*].

Communication: *Nick tells you that he lied about his opinion of the speech because he thought it would help you feel and perform better.*

Vignettes in Study 7

A. Healthcare vignette

Imagine that your spouse has a fatal cancer. You and your spouse met with the doctor, who informed you that your spouse's existing treatment has not been effective and that the cancer has spread to your spouse's bones and brain. You know your spouse may pass away soon and you have already prepared for the worst. However, the doctor says that there is always hope and that your spouse may qualify for a new experimental treatment soon. A few weeks later, your spouse passes away. You subsequently find out that the doctor knew that your spouse was too sick to receive any experimental treatments in the future. The doctor had known you and your spouse for around 6 months. He had never discussed your and your spouse's preferences for negative information. He did not know whether you and your spouse wanted to remain hopeful and optimistic, or whether you and your spouse wanted complete candor in such dire circumstances.

Ambiguous motivation: [*no additional information*].

Benevolent motivation: *In reality, the doctor lied about the experimental treatment options because he wanted to preserve your and your spouse's hope. He was sincerely trying to do what he thought was best for you and your spouse.*

B. Financial advisor vignette

Imagine that you are meeting with your financial advisor about potentially investing in a new fund. Investing in this fund would bring significant financial risk to you, but could also yield high rewards. You tell your financial advisor that you would like to invest in this fund. However, your advisor tells you that you do not meet the minimum criteria to invest. Several weeks later, you find out that you do in fact meet the criteria to invest in this fund, and that your financial advisor knew this. You and your financial advisor have known each other for around 6 months. You two had never discussed your desire to invest in high-risk/high-reward funds, or to stick with low-risk, low-reward funds.

Ambiguous motivation: [*no additional information*].

Benevolent motivation: *In reality, your advisor lied about you not meeting the criteria because s/he thought it would make you financially better off. S/he was sincerely doing what s/he thought was best for you.*

References

Bok, S. (1978). *Lying: Moral Choices in Public and Private Life.* New York: Pantheon.

Boles, T. L., Croson, R. T., & Murnighan, J. K. (2000). Deception and retribution in repeated ultimatum bargaining. *Organizational behavior and human decision processes*, *83*(2), 235-259.

Brehm, J. W. (1966). *A theory of psychological reactance*. New York: Academic Press.

Brehm, J. W., Stires, L. K., Sensenig, J., & Shaban, J. (1966). The attractiveness of an eliminated choice alternative. *Journal of Experimental Social Psychology*, *2*(3), 301-313.

Brockner, J., Konovsky, M., Cooper-Schneider, R., Folger, R., Martin, C., & Bies, R. J. (1994). Interactive effects of procedural justice and outcome negativity on victims and survivors of job loss. *Academy of Management Journal*, *37*(2), 397-409.

Brockner, J., & Wiesenfeld, B. M. (1996). An integrative framework for explaining reactions to decisions: interactive effects of outcomes and procedures. *Psychological Bulletin*, *120*(2), 189-208.

Croson, R., Boles, T., & Murnighan, J. K. (2003). Cheap talk in bargaining experiments: lying and threats in ultimatum games. *Journal of Economic Behavior & Organization*, *51*(2), 143-159.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353-380.

Cushman, F. (2013). Action, outcome, and value a dual-system framework for morality. *Personality and social psychology review*, *17*(3), 273-292.

deCharms, R. (1968). Personal causation. New York: Academic Press.

Deci, E. L., & Ryan, R. M. (1985). Intrinsic motivation and self-determination in human behavior. New York: Plenum.

Deci, E. L., & Ryan, R. M. (1987). The support of autonomy and the control of behavior. *Journal of Personality and Social Psychology*, *53*(6), 1024-1037.

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology*, *70*(5), 979-995.

Epley, N., Keysar, B., Van Boven L., & Gilovich, T. (2004). Perspective Taking as Egocentric Anchoring and Adjustment. *Journal of Personality and Social Psychology*, *87*(3), 327-339.

Erat, S., & Gneezy, U. (2012). White lies. *Management Science*, *58*(4), 723-733.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785-791.

Fitzsimons, G. J., & Lehmann, D. R. (2004). Reactance to recommendations: When unsolicited advice yields contrary responses. *Marketing Science*, *23*(1), 82-94.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, *117*(1), 21-38.

Gino, F., Ayal, S., & Ariely, D. (2013). Self-serving altruism? The lure of unethical actions that benefit others. *Journal of Economic Behavior and Organization*, *93*, 285-292.

Gino, F., & Pierce, L. (2009). Dishonesty in the name of equity. *Psychological Science*, *20*(9), 1153-1160.

Gino, F., Shu, L. L., & Bazerman, M. H. (2010). Nameless+ harmless = blameless: When seemingly irrelevant factors influence judgment of (un) ethical behavior. *Organizational Behavior and Human Decision Processes*, *111*(2), 93-101.

Gneezy, U. (2005). Deception: The role of consequences. *The American Economic Review*, *95*(1), 384-394.

Gneezy, U., Rockenback, B., & Serra-Garcia, M. (2013). Measuring lying aversion. *Journal of*

Economic Behavior & Organization, 93(C), 293-300.

Graham, J., Meindl, P., Koleva, S., Iyer, R., & Johnson, K. M. (2015). When values and behavior conflict: Moral pluralism and intrapersonal moral hypocrisy. *Social and Personality Psychology Compass*, *9*(3), 158-170.

Greenberg, A. E. (2016). Essays in behavioral economics.

Greenberg, A. E., Smeets, P., & Zhurakhovska, L. (2015). Promoting truthful communication through ex-post disclosure. Available at SSRN: http://ssrn.com/abstract=2544349.

Greenberg, A. E., & Wagner, A. F. (2016). The ripple effects of deceptive reporting. Available at SSRN: http://ssrn.com/abstract=2649392.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364-371.

Gunia, B. C., Wang, L., Huang, L., Wang, J., & Murnighan, J. K. (2012). Contemplation and conversation: Subtle influences on moral decision making. *Academy of Management, 55*(1), 13-33.

Halevy, N., & Chou, E. Y. (2014). How decisions happen: focal points and blind spots in interdependent decision making. *Journal of Personality and Social Psychology*, *106*(3), 398-417.

Halevy, N., & Halali, E. (2015). Selfish third parties act as peacemakers by transforming conflicts and promoting cooperation. *Proceedings of the National Academy of Sciences*, *112*(22), 6937-6942.

Hardisty, D. J., Thompson, K. F., Krantz, D. H., & Weber, E. U. (2013). How to measure time preferences: An experimental comparison of three methods. *Judgment and Decision Making*, *8*(3), 236-249.

Hayes, A. (2016). PROCESS macro for SPSS and SAS. Retrieved February 1, 2016 from http://www.processmacro.org/index.html.

Hsee, C. K., & Weber, E. U. (1997). A fundamental prediction error: Self–others discrepancies in risk preference. *Journal of experimental psychology: general*, *126*(1), 45-53.

John, L. K., Barasz, K., & Norton, M. I. (2016). Hiding personal information reveals the worst. *Proceedings of the National Academy of Sciences, 113*(4), 954-959.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of personality and social psychology*,*103*(1), 54-69.

Kant, I. (1785). Foundation of the metaphysics of morals. Beck LW, translator. Indianapolis: Bobbs-Merrill; 1959.

Levine, E.E. (2017). Community standards of deception. *Working paper.*

Levine, E. E., & Schweitzer, M. E. (2014). Are liars ethical? On the tension between benevolence and honesty. *Journal of Experimental Social Psychology*, *53*, 107-117.

Levine, E. E., & Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, *126*, 88-106.

Lupoli, M. J., Jampol, L. E., & Oveis, C. (2017). Lying because we care: Compassion increases prosocial lying. *Journal of Experimental Psychology: General*, *146*(7), 1026-1042.

McFarlin, D. B., & Sweeney, P. D. (1992). Distributive and procedural justice as predictors of satisfaction with personal and organizational outcomes. *Academy of Management Journal*, *35*(3), 626-637.

Miller, R. M., Hannikainen, I. A., & Cushman, F. A. (2014). Bad actions or bad outcomes? Differentiating affective contributions to the moral condemnation of harm. *Emotion*, *14*(3), 573-587.

Miron, A. M., & Brehm, J. W. (2006). Reactance theory-40 years later. *Zeitschrift für Sozialpsychologie*, *37*(1), 9-18.

Murnighan, J. K., & Wang, L. (2016). The social world as an experimental game. *Organizational Behavior and Human Decision Processes, 136*(C), 89-94.

Nakashima, E. (2016). Apple vows to resist FBI demand to crack iPhone linked to San Bernardino attacks. *The Washington Post*. Retrieved March 16, 2015 from https://www.washingtonpost.com/world/national-security/us-wants-apple-to-help-unlock-iphone-used-by-san-bernardino-shooter/2016/02/16/69b903ee-d4d9-11e5-9823-02b905009f99_story.html

Nisbett, R., & Wilson, T. (1977). Telling more than we know: verbal reports on mental processes. *Psychological Review, 84*(3), 231-259.

Ostrom, E., Gardner, R., & Walker, J. (1994). *Rules, Games, and Common-Pool Resources.* Ann Arbor: The University of Michigan Press.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior research methods, instruments, & computers*, *36*(4), 717-731.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*(3), 879-891.

Rapoport, A. (1973). *Two-person game theory*. Courier Corporation.

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD Triad Hypothesis: A Mapping Between Three Moral Emotions (Contempt, Anger, Disgust) and Three Moral Codes (Community, Autonomy, Divinity). *Journal of Personality and Social Psychology*, *76*(4), 574-586.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68-78.

Schweitzer, M. E., & Croson, R. (1999). Curtailing deception: The impact of direct questions on lies and omissions. *International Journal of Conflict Management*, *10*(3), 225-248.

Schweitzer, M. E., Hershey, J. C., & Bradlow, E. T. (2006). Promises and lies: Restoring violated trust. *Organizational behavior and human decision processes*, *101*(1), 1-19.

Shalvi, S., Dana, J., Handgraaf, M. J., & De Dreu, C. K. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, *115*(2), 181-190.

Shalvi, S., Gino, F., Barkan, R., & Ayal, S. (2015). Self-serving justifications doing wrong and feeling moral. *Current Directions in Psychological Science*, *24*(2), 125-130.

Shu, L. L., Gino, F., & Bazerman, M. H. (2011). Dishonest deed, clear conscience: When cheating leads to moral disengagement and motivated forgetting. *Personality and Social Psychology Bulletin*, *37*(3), 330-349.

Shweder, R., Much, N., Mahapatra, M., & Park, L. (1997). Divinity and the "big three" explanations of suffering. *Morality and Health*, *119*, 119-169.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, *89*(6), 845-851.

Sutter, M. (2009). Deception through telling the truth?! Experimental evidence from individuals and teams. *Economic Journal*, *119*(534), 47-60.

Tannenbaum, D., Fox, C. R., & Rogers, T. (2016). On the misplaced politics of behavioral policy interventions. *Working paper*.

Tannenbaum, D., & Ditto, P. H. (2016). Information Asymmetries in Default Options. *Working paper*.

Tyler, T. R., Degoey, P., & Smith, H. J. (1996). Understanding why the justice of group procedures matters: A test of the psychological dynamics of the group-value model. *Journal of Personality and Social Psychology, 70*, 913–930.

Tyler, J., Feldman, F., & Reichert, A. (2006). The price of deceptive behavior: Disliking and lying to people who lie to us. *Journal of Experimental Psychology, 42*(1), 69-77.

Van Boven, L., & Loewenstein, G. (2003). Social projection of transient drive states. *Personality and Social Psychology Bulletin*, *29*(9), 1159-1168.

Wiltermuth, S. S. (2011). Cheating more when the spoils are split. *Organizational Behavior and Human Decision Processes*, *115*(2), 157-168.

Zhong, C. B. (2011). The ethical dangers of deliberative decision making. *Administrative Science Quarterly*, *56*(1), 1-25.