Creating Artificial Intelligence: An Inductive Study of How Creative Workers Forecast the Future and Manage Present Emotions

Author: Lydia Paine Hagtvedt

CREATING ARTIFICIAL INTELLIGENCE: AN INDUCTIVE STUDY OF HOW CREATIVE WORKERS FORECAST THE FUTURE AND MANAGE PRESENT EMOTIONS

by

Lydia Paine Hagtvedt

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Management and Organization

Boston College Carroll School of Management

Doctoral Committee:
Spencer H. Harrison (Co-Chair), Boston College
Michael G. Pratt (Co-Chair), Boston College
Teresa M. Amabile, Harvard Business School

ABSTRACT

Through an inductive, qualitative study of individuals developing new artificial intelligence (AI) technologies, this dissertation builds theory on how creative workers manage the emotions that arise from forecasting the outcomes of implementing their creations. I find that, in a context that illuminates the danger of implementing certain types of creative ideas, creative workers forecast both positive and negative outcomes arising from implementing their work, which elicits ambivalence. My work indicates that how creative workers respond to this ambivalence affects whether they impose constraints on their work as it unfolds. First, some individuals may proceed without constraints because they have resolved their ambivalence by amplifying their positive thoughts and feelings toward their work. Informants who exhibited this pattern created psychological distance (Lewin, 1951; Trope & Liberman, 2003) from the potential negative effects of their work by anchoring on the present moment and/or emphasizing potential positive outcomes. However, the majority of informants exhibited a novel "redistribution" response to ambivalence, whereby they committed to their work (Brickman et al., 1987; Pratt & Rosa, 2003; Pratt & Pradies, 2011) and shifted from a strengthening of negative thoughts and feelings toward a strengthening of positive thoughts and feelings through the use of self-imposed constraints. My work suggests that, although self-imposed constraints do not eliminate negative thoughts and feelings altogether, applying these self-determined boundaries enables individuals to reduce ambivalence and engage (Harter, Schmidt, & Hayes, 2002) more fully in their work. In addition to inducing a process model that encompasses these dynamics, I present the categories and types of self-imposed constraints that I have induced. These self-imposed constraints are not mutually exclusive, and each serves one of three broader purposes: developing a sense that one's creation will have a positive moral valence, that one will

be able to control his or her creation, or that one may trust in the quality of his or her creation. This dissertation extends theory on the role of prospective thought processes in creative work and shows how constraints, though often seen as impediments to creativity, can be used proactively by creative workers to manage the darker emotions and thought processes that have largely been overlooked in prior research. This work also contributes a novel response to ambivalence, redistribution, which entails approaching potentially harmful creative work in a heedful manner.

ACKNOWLEDGEMENTS

I thank my committee; Spencer, Mike, and Teresa; my colleagues; and my family for pushing me in my thinking and encouraging me throughout the dissertation process. When not immersed in conducting interviews and analyzing the ideas that were emerging from my analysis, I have been diligently filing away the plethora of emails from colleagues, family, and friends with subject lines taking some form of: “More AI stuff.” In the final phase of my dissertation process, I have been very grateful for everyone who took the time to send resources my way. These digital reminders of the big picture prompted me to forecast how my own work could contribute to the scholarly understanding of creativity and possibly even to the future of AI.

TABLE OF CONTENTS

CHAPTER 1:
INTRODUCTION

> So much of the community is in a phase of just getting these things to work, that we're just starting to address the problem of, "Well, what if they work *too well*?" And, "Even if they work perfectly well, what do we use them for?" And I think you've caught me at a time where I am not entirely sure about that, and I'm thinking about it a lot. [Harry]

Creativity, defined as the production of novel and useful ideas or products (Amabile, 1983, 1988; Oldham & Cummings, 1996; Woodman, Sawyer, & Griffin, 1993; Stein, 1974), is part of what makes us human and enables us to flourish. But inherent in the definition of creativity, specifically the notion of novelty, is the possibility that negative outcomes may arise when introducing a creative product to the world. It is impossible to know ahead of time all of the unintended and possibly negative effects that might result from implementing a creative idea. To illustrate, McDonald's revolutionized the food industry by providing consumers with convenient and inexpensive meals, leading many other fast food franchises to follow suit – and driving the ongoing obesity epidemic (e.g., Bowman & Vinyard, 2004; Pereira, Kartashov, Ebbeling, Van Horn, Slattery, Jacobs, & Ludwig, 2005). As another example, the invention of the smartphone equipped people with pocket-sized computers, enabling the near-constant accessibility of information and social connections – and initiating a smartphone addiction crisis (Demirci, Akgönül, & Akpinar, 2015). As a final example, e-cigarettes replace tobacco with electronically heated nicotine, helping smokers to quit (Hajek et al., 2019) – and sparking nicotine addiction and a potentially fatal lung illness among younger generations (Tolentino, 2018; Abbott, 2019). Creative inventions have the potential to effect significant change when they are implemented, but positive outcomes may be accompanied by adverse effects, and some ideas may eventually do more harm than good.

Research has held that creativity relies on being able to explore ideas freely rather than being overly attuned to outcomes (e.g., Amabile, 1979). However, the uncertainty inherent in implementing truly novel creations likely compels individuals to forecast the downstream effects that could arise from implementing their ideas.[1] Some research has begun to explore the prospective thought process of creative forecasting, which is defined as predicting the outcomes of new ideas (Kettner, Guilford, & Christensen, 1959; Wilson, Guilford, Christensen, & Lewis, 1954; Byrne, Shipman, & Mumford, 2010; Berg, 2016). This work has assessed individuals' accuracy in predicting whether an idea is likely to succeed (e.g., Mumford, Lonergan, & Scott, 2002; Byrne, Shipman, & Mumford, 2010; Berg, 2016) and explored how individuals may revise their ideas based on forecasting whether they will meet implementation criteria (e.g., Lonergan, Scott, & Mumford, 2004). However, extant research largely neglects the forecasting that creative workers may do regarding what may happen *after* an idea becomes highly successful. Indeed, the more widely adopted an idea, the greater its potential to do harm, should unintended consequences arise.

Because creativity involves developing ideas that challenge existing norms (Brockhaus, 1980; Janssen & Van Yperen, 2004; Kahneman & Tversky, 1979; Amabile, 1983, 1988; Amabile & Pratt, 2016; Mumford & Gustafson, 1988; Shalley, Zhou, & Oldham, 2004; Zhou & George, 2001; Perry-Smith, 2006; Simmons & Ren, 2009), implementing creative ideas may bring about unintended – and perhaps dangerous – outcomes. Some creative ideas, such as complex, new technologies, have a heightened potential for harmful effects on large numbers of people due to their potential scope of impact and the ways that they are integrated into people's

[1] Throughout this dissertation, I use phrases like "positive and negative effects," "potential outcomes", "distal outcomes," "forecasted outcomes," "potential consequences," "potential adverse effects," and "downstream effects" to refer to the outcomes that follow the implementation of creative ideas.

lives (Jasper, 2010). I refer to this as "potentially harmful creativity." With growing discussion of the likelihood of adverse downstream effects on the general public, the domain of artificial intelligence (AI) development, which involves computer systems that rival or surpass human intelligence in their capabilities, exemplifies these dynamics. Under such circumstances, the creative workers developing the new technologies may grapple with the negative outcomes that could follow the implementation of their creations. This dissertation examines this tension, exploring how creative workers may persist despite the potential dangers posed by implementing their ideas.

One might think that individuals simply would not engage in potentially harmful creativity, yet retrospective accounts (e.g., Bethe, 1968; Feynman, 2005) indicate that individuals do indeed pursue creative ideas even when there may be negative effects on others (Hecht, 2010; Zaitseva, 2010; Cropley, 2010). Hecht (2010) speculates that the scientists involved in developing nuclear technology, for example, willfully blinded themselves to the possible consequences of their work. However, there is a paucity of research examining potentially harmful creativity as it unfolds, and in particular, the cognitive, emotional, and behavioral reactions that may be elicited by forecasting downstream effects. Considering downstream effects may impact the decisions that creative workers make during the creative process by evoking emotional reactions that influence thinking and behavior. For instance, sensing that one is negatively affecting other people can give rise to guilt (e.g., Baumeister, Stillwell, & Heatherton, 1994), and expecting negative outcomes is associated with emotions and thought processes that drive more cautious behavior (e.g., Sitkin & Weingart, 1995). At the same time, creativity depends on how individuals feel (e.g., positive affect, Amabile, Barsade, Mueller, & Staw, 2005), think (e.g., creativity-relevant processes, Amabile, 1983, 1988, 1996; Amabile &

Pratt, 2016), and act (e.g., challenging the status quo, Zhou & George, 2003). However, forecasting potential downstream effects has largely been overlooked in models seeking to explain key influences on creative outcomes (e.g., Amabile & Pratt, 2016; see Lubart, 2001, for a review).

This absence is likely an artifact of several factors. First, there is a longstanding divide between creativity and innovation research. Creativity research primarily focuses on the initial generation of creative ideas and products, while research on innovation, defined as the implementation of creative ideas at the organization or field level, generally focuses on these "macro-level" dynamics (Amabile, 1988; West & Farr, 1990; Amabile & Pratt, 2016; Hennessey & Amabile, 2010; George, 2007; Mumford et al., 2002; Byrne et al., 2010). In reality, this split is somewhat artificial; creative workers often do think about whether their ideas will be implemented (e.g., Berg, 2016). Second, and related to the divide between creativity and innovation research, creativity is often studied through laboratory experiments (Hennessey & Amabile, 2010; Elsbach & Kramer, 2003; Drazin, Glynn, & Kazanjian, 1999; Oldham & Cummings, 1996), which by nature rely on short-term, finite tasks. Such tasks enable the precise evaluation of causal influences on creative outcomes, but they typically preclude examining factors associated with implementation, even though considerations regarding the consequences of implementation may play an important role in the creative process.

Third, because creativity and innovation are frequently assumed to be inherently beneficial for society (James, Clark, & Cropanzano, 1999; James & Taylor, 2010, Runco, 2010), extant research has not taken seriously the notion that creative ideas can be dangerous (Anderson, Potočnik, & Zhou, 2014). Prior research on potentially harmful creativity is quite limited and highly speculative. It relies primarily on descriptive, retrospective accounts, such as those of the scientists who developed nuclear technology or who participated in the Soviet Union's program for developing weapons of mass destruction (e.g., Bethe, 1968; Zaitseva, 2010). Cropley (2010) posits that despite having good intentions, individuals engaging in potentially harmful creativity may be unable or unwilling to consider the potential consequences of their work, whether because of their own intrinsic enjoyment or because of external coercion. Whether either of these explanations (among others) is accurate remains in question, though, due to the reliance on retrospective accounts. Such accounts are colored by hindsight bias (Roese & Vohs, 2012) and therefore neglect important sensemaking processes that occur along the way to realizing creative outcomes (Drazin, Glynn, & Kazanjian, 1999). Indeed, creative workers must navigate the potential outcomes of implementing their ideas in the moment, when the novelty of the emerging idea occludes the exact nature of possible outcomes, even though they may be broadly aware that downsides will likely exist. This is a significant distinction, because post hoc justifications differ from in-the-moment reasoning due to the innate equivocality and emotional charge of in-the-moment interpretations and judgments (March, 1994; Weick, 1979, 1995; Hogarth, Portell, & Cuxart, 2007; Hogarth, Portell, Cuxart, & Kolev, 2011).

These shortcomings create an opportunity to extend the understanding of the role of prospective thought processes in creative work by exploring whether and how forecasting distal outcomes elicits emotional reactions and influences how creative workers approach their work in the present. Because of the greater likelihood of adverse effects on others, engaging in potentially harmful creativity may elicit particularly strong emotional reactions as individuals consider the possible outcomes of their work. Though seeing the positive potential of a creation may drive excitement and enjoyment (Scheier & Carver, 1985, 1992; Andersson, 1996), forecasting negative effects may engender concern and guilt (e.g., Baumeister, Stillwell, &

Heatherton, 1994), which may change how creative workers approach their work. Another possibility is that positive and negative outcomes may seem equally likely, which could be associated with emotional ambivalence (positive and negative feelings, e.g., Pratt & Doucet, 2000; Fong, 2006; Fong & Tiedens, 2002; Larsen, McGraw, & Cacioppo, 2001; Williams & Aaker, 2002; Rothman & Wiesenfeld, 2007).

With this dissertation, I examine the following research questions regarding forecasting and emotions in the creative process:

> ***Research Question 1a:*** How, if at all, do creative workers forecast the outcomes of implementing their potentially harmful creations?
> ***Research Question 1b:*** What emotional experiences, if any, arise during this forecasting process?

In addition to unpacking the relationship between creative forecasting and the emotions it engenders, I also investigate whether and how this influences the creative process:

> ***Research Question 2:*** How, if at all, does forecasting the outcomes of implementing potentially harmful creations influence the creative process?

The growing domain of artificial intelligence (AI), which involves computer systems that rival or surpass human intelligence in their capabilities, offers a fruitful context to explore these dynamics with inductive, qualitative methods. With AI, the nature of what is created greatly amplifies the potential for adverse effects on a large number of people. Though AI may solve many of society's problems, from biased decision making to car accidents to disease (e.g., Vyas, 2018; Tegmark, 2019), AI promises to displace many human workers, and it threatens societal norms and even human safety, if not properly applied (Tegmark, 2019). Given the threats to the general public posed by AI development and implementation, these dynamics create a rich context to explore how creative workers forecast the outcomes of implementing their work, the emotions that arise, and how these factors influence the creative process.

I employ inductive, qualitative methods to address my research questions. To preview my findings, my work indicates that, in the context of potentially harmful creativity, creative workers forecast both positive and negative outcomes of implementing their work, which elicits ambivalence. How individuals respond to this ambivalence seems to affect whether they impose constraints on their work as it unfolds. One group of informants amplified their positive thoughts and feelings toward their work and proceeded without constraints. These informants created psychological distance (Lewin, 1951; Trope & Liberman, 2003) from the potential negative effects of their work by anchoring on the present moment and/or emphasizing the relative importance of potential positive outcomes. They thereby rendered potential negative outcomes irrelevant to their work and did not constrain their work. However, the majority of informants exhibited a novel "redistribution" response to ambivalence, whereby they committed to their work (Brickman et al., 1987; Pratt & Rosa, 2003; Pratt & Pradies, 2011) but shifted from a temporary intensification of negative thoughts and feelings toward a strengthening of positive thoughts and feelings through the use of self-imposed constraints. My work suggests that, although self-imposed constraints do not eliminate negative thoughts and feelings altogether, applying these self-determined boundaries enables individuals to take responsibility for potential negative outcomes and quell fears about the future of their creations. This process ultimately seems to enable them to engage (Harter, Schmidt, & Hayes, 2002) more fully in their work.

This dissertation makes important contributions to theory on forecasting, emotions, and constraints in creative work. Although the creative process is often suffused with positive emotion (e.g., Isen, 2000; Isen & Reeve, 2005; Amabile, Barsade, Mueller, & Staw, 2005; Amabile & Kramer, 2011), and creativity itself is lauded as a benefit to society (James, Clark, & Cropanzano, 1999; James & Taylor, 2010), creative workers still have to grapple with complicated and often negative thoughts and emotions about what their creations could become. I show how, in potentially harmful creative work, ambivalence accompanies forecasting distal outcomes. I reveal how different responses to ambivalence inform the creative process and may reshape its outcomes, depending on whether individuals impose constraints on their work and the types of constraints that they use. This dissertation also helps build a bridge between research on creativity and innovation by showing how prospective thought processes regarding implementation, typically a concern of innovation, may influence the choices that individuals make during the creative process. Additionally, this work answers the recent call for organizational scholars to examine AI development in particular – specifically, how individuals developing new AI technologies think about the downstream effects of their work and how these considerations affect the decisions that they make in the present (Amabile, 2019).

Prior work on creativity and emotions has focused largely on emotions as an input into the creative process, showing how positive affect seems to be a consistent driver of creativity (see Hennessey & Amabile, 2010, and Amabile & Pratt, 2016), while negative affect may contribute to creative performance under particular circumstances (e.g., George & Zhou, 2002; George & Zhou, 2007). Rather than viewing emotions as an input into the creative process or focusing on emotions of one valence, I show how mixed emotions may emerge *during* the creative process, when individuals forecast both positive and negative outcomes of implementing their work, which seems to influence their work directly.

I also build theory by showing how constraints (e.g., Rosso, 2014; Amabile & Gitomer, 1984) can be used to help creative workers manage the darker thoughts and emotions that have largely been overlooked in prior research seeking to explain the creative process (e.g., Amabile & Pratt, 2016; see Lubart, 2001, for a review). To my knowledge, prior research on constraints has focused exclusively on external constraints and has often shown how they inhibit creativity (e.g., Woodman, Sawyer, & Griffin, 1993; Amabile, 1983, 1988, 1996), though some work has indicated that certain types of external constraints may benefit idea generation (e.g., Finke, 1990). I offer a different perspective, revealing how constraints may be used proactively (i.e., self-imposed) by creative workers to manage the cognitive and emotional challenges posed by engaging in potentially harmful creativity.

From a managerial perspective, this work sheds light on how organizations might best manage potentially harmful creativity. This project reveals that in the absence of field-level norms and ethical guidelines to guide behavior, individuals working on potentially harmful creative endeavors respond differently to the ambivalence that arises from forecasting the outcomes of implementing their work, and these differences appear to influence whether they choose to constrain their own creativity. Organizations involved in potentially harmful creativity should recognize, first of all, that ambivalence is likely prevalent among employees, and how individuals respond may reshape the creative process as well as the ideas that they develop. Organizations should work to understand employees' emotional experiences and determine whether their constraint decisions align with organizational goals. Further, this dissertation contributes to an emerging discussion on the consequences of unconstrained creativity, whether for individuals' personal lives (e.g., Harrison & Wagner, 2016) or for society at large (Cropley, 2010). I demonstrate how self-imposed constraints may help individuals to engage heedfully in potentially harmful creative work.

Chapter 2 presents the literature review for this dissertation, focusing on creativity and the risk of harmful outcomes, forecasting, and emotions. Chapter 3 details the inductive, qualitative methodological approach and rationale for the study. Chapter 4 presents the findings, including the induced process model illustrating forecasting, ambivalence responses, and the role of constraints in the creative process, as well as the induced categorization of the different types of constraints that informants imposed on their work. Finally, in Chapter 5, the dissertation closes with a discussion of theoretical contributions to creativity research and areas for future research.

CHAPTER 2:
LITERATURE REVIEW

Creativity, defined as the production of novel and useful ideas or products (Amabile, 1983, 1988; Oldham & Cummings, 1996; Woodman, Sawyer, & Griffin, 1993; Stein, 1974), is often assumed to be a purely positive force (Runco, 2010; James, Clark, & Cropanzano, 1999; James & Taylor, 2010; Anderson, Potočnik, & Zhou, 2014). However, novelty carries the risk of unintended – and possibly negative – outcomes. Particularly when creativity is potentially harmful, defined as having the potential to adversely affect a large number of people through implementation (application of idea at organization or field level; Amabile, 1988; West & Farr, 1990), workers may grapple with the possible outcomes of implementing their work. This may give rise to emotions that influence the creative process as it unfolds. However, existing models of the creative process (e.g., Amabile, 1983, 1988; Amabile & Pratt, 2016; see Lubart, 2001, for a review) largely assume that creative workers strive for successful and impactful ideas, overlooking how forecasting downstream effects may elicit complicated and possibly negative emotions. As such, existing models neglect a potentially significant influence on the process by which individuals develop their ideas.

In the sections that follow, I review relevant research on creativity and the risk of harmful outcomes, creative forecasting, and emotions. I argue that the process of forecasting distal outcomes during potentially harmful creative endeavors demands further scholarly attention.

Creativity, risk, and the potential for harm. Creativity involves taking the risk of proposing a new idea: By definition, ideas are only creative if they are novel (Amabile, 1983, 1988; Oldham & Cummings, 1996; Woodman, Sawyer, & Griffin, 1993; Stein, 1974). A large body of work relates creativity to being willing to push boundaries and pursue ideas that break with existing norms (Brockhaus, 1980; Janssen & Van Yperen, 2004; Kahneman & Tversky, 1979; Amabile, 1983, 1988; Amabile & Pratt, 2016; Mumford & Gustafson, 1988; Shalley, Zhou, & Oldham, 2004; Zhou & George, 2001; Perry-Smith, 2006; Simmons & Ren, 2009). Because creativity inherently entails exploring new frontiers, either incrementally or more radically (Mumford & Gustafson, 1988), strategies for fostering creativity typically involve questioning the status quo, breaking rules, and taking risks (Baucus, Norton, Baucus, & Human, 2008). Indeed, risk aversion has a negative effect on creativity (Friedman & Förster, 2001) because focusing on risks prevents individuals from considering more novel responses.

Embedded in the notion of novelty, however, is the risk of unintended and possibly negative outcomes. Further, some types of creative ideas have a particularly wide scope of impact associated with their implementation – and a heightened potential to adversely affect large numbers of people. I refer to this as "potentially harmful creativity." This potential for widespread harm is especially pronounced in fields that involve the development of complex new technologies (Jasper, 2010), such as nuclear technology, gene editing, and artificial intelligence (AI; computer systems that rival or surpass human intelligence). Nonetheless, precisely because of the potential scope of impact, such creativity may also offer great benefits (e.g., Hecht, 2010). As an example, the same creative ideas that led to the development of nuclear bombs, which killed an estimated 246,000 people in Japan at the end of World War II, also enabled the creation of nuclear energy, which has saved an estimated 1.84 million lives by reducing carbon pollution (Kharecha & Hansen, 2013).

Although we know that creative workers must be able to take risks in terms of pursuing novel ideas and that certain types of ideas have a heightened potential for harmful outcomes, we know relatively little about how creative workers actually forecast the distal risks associated with the implementation of their ideas. The absence of research on how creative workers forecast the potential outcomes of implementation is likely due to several factors.

First, there is a longstanding split between the creativity and innovation literatures, with creativity research focusing on individual or group idea generation and innovation research focusing on the macro-level dynamics of innovation, defined as the implementation of new ideas at the organization or field level (Amabile, 1988; West & Farr, 1990; Amabile & Pratt, 2016; Hennessey & Amabile, 2010; George, 2007). Embedded in this divide is the assumption that creative workers do not concern themselves very much with implementation (see *Creative forecasting* section below for exceptions). Second, experimental research, which makes up the bulk of creativity research (Hennessey & Amabile, 2010; Elsbach & Kramer, 2003; Drazin, Glynn, & Kazanjian, 1999; Oldham & Cummings, 1996), typically relies on short-term, finite tasks. While some laboratory experiments have asked individuals to envision particular implementations of their ideas (e.g., Lonergan, Scott, & Mumford, 2004; Byrne, Shipman, & Mumford, 2010), it is unclear whether conclusions based on such designs would hold if individuals are truly invested in the outcomes of implementation, as they typically are in organizational settings.

A third issue is the frequent – and often untested – assumption that creativity and innovation are purely positive forces. The potential dangers of developing and implementing certain types of creative ideas are generally overlooked (Anderson, Potočnik, & Zhou, 2014). To my knowledge, the only scholarly work that has attempted to shed light on the dynamics of potentially harmful creativity consists of post-hoc analyses of retrospective accounts (e.g., Bethe, 1968). As an example, by analyzing retrospective reports from the scientists who developed nuclear technology, scholars have posited that scientists proceeded despite the risk of significant adverse effects due to an intentional blindness to possible negative consequences (Hecht, 2010), a "fascination or unquestioning belief in what they [were] doing" (Cropley, 2010: 360), or simply "complacency or hubris" (Jasper, 2010: 111). Retrospective accounts from others involved (e.g., Feynman, 2005) introduce the possibility that at least some of the scientists believed that dangerous consequences were likely but proceeded because they reasoned that it would be more dangerous *not* to make advancements. As another example, based on post-hoc and often secondary accounts (e.g., Wheelis, Rózsa, & Dando, 2006) derived from a number of sources, Zaitseva (2010) argues that the scientists who developed biological weapons for the Soviet Union were motivated by prestige, the ability to participate in creative work with unrestricted resources, and the desire to protect their home country.

Of course, retrospective accounts differ significantly from in-the-moment descriptions. In-the-moment reasoning is by nature more equivocal and uncertain (March, 1994; Weick, 1979, 1995; Sonenshein, 2007). Retrospective accounts are subject to hindsight bias, which leads individuals to act as though previous events were expected, when in reality, they could not have been predicted (Roese & Vohs, 2012). This "creeping determinism" (Nestler, Blank, & von Collani, 2008: 182) often prevents individuals from accounting for in-the-moment thought processes and actions after the fact. In-the-moment perceptions also tend to be more emotionally charged than retrospective accounts of the same issues (Hogarth, Portell, & Cuxart, 2007; Hogarth, Portell, Cuxart, & Kolev, 2011), and given that emotions serve as a source of information (Schwarz & Clore, 1983; Schwarz, 2002; Schwarz & Clore, 2003; Schwarz, 2011), the dulled emotions of retrospective accounts (e.g., Aaker, Drolet, & Griffin, 2008) may not adequately explain in-the-moment decisions. Conclusions based on retrospective accounts are therefore speculative and potentially incorrect. It is critical to understand the prospective thought processes that inform creativity as it unfolds, before potential hazards become concrete outcomes and before in-the-moment interpretations are replaced by retrospective justifications.

Below, I discuss existing research on creative forecasting before arguing that the paucity of research in this area demands further scholarly attention due to how the outcomes of forecasting may influence the creative process.

Creative forecasting. Creative forecasting is defined as predicting the outcomes of new ideas within particular settings (Kettner, Guilford, & Christensen, 1959; Wilson, Guilford, Christensen, & Lewis, 1954; Byrne, Shipman, & Mumford, 2010; Berg, 2016). Forecasting has been conceptualized as part of the "late cycle process" (Byrne et al., 2010: 119) of idea evaluation, during which individuals appraise their ideas based on a particular set of standards (Mumford, Lonergan, & Scott, 2002; Lonergan, Scott, & Mumford, 2004). To delineate how creative forecasting fits into predominate stage models of the individual creative process (e.g., Wallas, 1926; Hogarth, 1980; Nystrom, 1979; Amabile, 1983, 1988), I begin by describing the stages that appear in most existing models and situate forecasting within the idea evaluation stage.

Situating creative forecasting in the creative process. Creativity begins with the task presentation stage (Amabile, 1983, 1988; Amabile & Pratt, 2016), sometimes called problem or task identification (Lubart, 2001), which is followed by preparation (Amabile, 1983, 1988; Amabile & Pratt, 2016), during which relevant resources are gathered. This may be accompanied by an incubation period (Gardner, 1993; Gruber & Davis, 1995; Mainemelis, 2010). Idea (response) generation follows, during which creative workers generate a number of possible solutions for the task at hand (Amabile, 1983, 1988; Lubart, 2001; Amabile & Pratt, 2016). The early stages of the creative process focus heavily on divergent thinking (Guilford, 1967, 1968, 1982), which involves making associations between diverse concepts and searching for original ideas. In the later stages, individuals utilize more convergent thinking, relying on experience and expertise to narrow down the search and focus on achieving creative success with a single, more developed idea (Cropley, 2006).

After an idea is (tentatively) selected, individuals engage in the "late cycle" (i.e., late-stage; Byrne et al., 2010: 119) processes of idea evaluation (validation), which involves checking the idea against the task criteria (Amabile, 1983, 1988; Amabile & Pratt, 2016) and possibly revising ideas based on the criteria (Mumford et al., 2002; Lonergan et al., 2004), and idea elaboration, which entails refining and adding detail to ideas as individuals prepare for implementation (Csikszentmihalyi, 1997; Mainemelis, 2010).[2] Because forecasting often entails predicting the likelihood of successful implementation, forecasting is an important component of idea evaluation (Mumford et al., 2002). It is also an essential precursor to the final stage of outcome assessment, which involves deciding whether to move forward with implementing an idea or to return to earlier stages to make changes (Amabile, 1983, 1988; Amabile & Pratt, 2016).

Creative forecasting. The concept of forecasting (defined as predicting outcomes more broadly) originates in psychology, and foundational work has focused on individuals' ability to make accurate predictions about the future (e.g., Pant & Starbuck, 1990; Mumford, Schultz, & Van Doorn, 2001; Thomas, Clark, & Gioia, 1993; Önkal, Yates, Sigma-Mugan, & Öztin, 2003). Due to the longstanding divide between research on creativity, which has focused largely on the generation of creative ideas, and innovation, which focuses on their implementation (Amabile,

[2] Once individuals select a "final" idea, the idea may then be evaluated by external figures, such as domain experts (e.g., Amabile, 1982). However, the focus of the present discussion is the individual's evaluation of his or her own ideas as candidates for further development.

1988; West & Farr, 1990; Amabile & Pratt, 2016; Hennessey & Amabile, 2010; George, 2007; Mumford et al., 2002; Byrne et al., 2010), research on creative forecasting is relatively limited (Mumford, 2001; Mumford et al., 2002; Byrne et al., 2010; Berg, 2016).

Within creativity research, the focus on prediction accuracy has manifested in research on individuals' ability to judge the novelty (i.e., originality; Runco & Smith, 1992; Basadur, Runco, & Vega, 2000; Licuanan, Dailey, & Mumford, 2007; Silvia, 2008) or market success (Berg, 2016) of their ideas. For instance, Runco and Smith (1992) found that individuals were better at assessing the originality of their own ideas than at assessing their popularity, and they provide evidence that assessing originality is distinct from other creative thinking skills. Licuanan and colleagues (2007) found that the tendency to underestimate the originality of highly novel ideas may be reduced by making novelty evaluation an active process and by making interactional processes more salient. Additionally, Silvia (2008) found that the skill of novelty evaluation is enhanced by openness to experience. Through a study of circus arts professionals and an accompanying lab experiment, Berg (2016) built on this prior work and examined predictions about market success. He found that creative workers were better than managers at forecasting the success of others' new ideas – but not their own. This (qualified) advantage appeared to be due to the combination of divergent and convergent thinking required for creativity as compared with the emphasis on convergent thinking for the managerial role. By focusing on individuals' ability to forecast the originality and success of new ideas, this body of work positions forecasting as a skill that increases the likelihood of creative success.

Yet the scholarly understanding of creative forecasting has begun to reach beyond examining whether individuals accurately forecast the originality and success of particular ideas. Some work suggests that the idea evaluation stage, which includes forecasting, may be an important influence on creative ideas themselves. For instance, in a study of marketing and R&D managers, Frankwick and colleagues (1994) found that individuals revised their strategic decision regarding developing a new technology as their understanding of what would drive organizational success changed. Forecasting may also affect whether individuals enhance their ideas based on the nature of the forecasted implementation. In a study of undergraduates proposing advertising campaigns, Lonergan and colleagues (2004) found that the creativity of ideas depended on the originality of the initial idea as well as the focus of implementation. Superior ideas resulted when novel initial ideas were combined with implementation goals emphasizing the efficiency of current processes *or* when less-novel initial ideas were paired with implementation goals emphasizing changing to something new. Further, the extent of forecasting may itself improve ideas. In another study of undergraduates proposing advertising campaigns, Byrne and colleagues (2010) found that more extensive forecasting boosted the quality, originality, and elegance of ideas.

These streams of research show that creative forecasting is not only a process of predicting whether ideas will be original and succeed. Forecasting may also involve predicting how implementation will unfold, and forecasting whether implementation will be successful may prompt individuals to reshape their ideas to increase the likelihood of success. However, what remains, to my knowledge, unexplored is how individuals forecast the potential downstream effects of their work – how they consider the outcomes that may *follow* generating and implementing a successful idea. This may be an important influence on the creative process in its own right. As I argue below, forecasting the aftermath of implementation may influence the creative process due to the emotions that are engendered by considering possible downstream effects.

Creative forecasting and emotions. Research has, to my knowledge, neglected an important likely byproduct of forecasting: the emotional reactions that may arise when thinking through the outcomes that may follow the successful implementation of a creative idea. The emotions that emerge from forecasting such distal effects warrant scholarly examination because individuals often experience powerful emotions when imagining future scenarios (Taylor & Schneider, 1989; Taylor, Pham, Rivkin, & Armor, 1998), and creative work is particularly vulnerable to the influences of affective states (Kaufmann, 2003; Hennessey & Amabile, 2010; Amabile & Pratt, 2016).

With the "affective revolution" in organizational behavior (Amabile & Pratt, 2016: 173), a large body of work has explored the influence of affect on creativity. Much of the extant research on affect and creativity comes from experimental or observational quantitative studies that examine positive or negative affect as an input into the creative process and investigate resulting creative performance. This research indicates that positive affect drives creativity more reliably than negative affect (Hennessey & Amabile, 2010; Amabile & Pratt, 2016). For instance, positive affect fosters intrinsic motivation (e.g., Isen & Reeve, 2005) and enhances creative thinking processes, such as by enabling access to a wider array of associations (e.g., Aspinwall, 1998; Isen, 2000; Fredrickson, 2001). Some field work has provided further evidence for the link between positive affect and creativity (Amabile, Barsade, Mueller, & Staw, 2005; Amabile & Kramer, 2011; Madjar, Oldham, & Pratt, 2002). For instance, Amabile and colleagues' (2005) field study of professionals in chemicals, high tech, and consumer products companies found that positive affect both positively influenced creative work and was the most frequent affective response to creative episodes. Additionally, Amabile and Kramer's (2011) work indicates that making even a small amount of progress in meaningful work drives positive affect as well as intrinsic motivation, which is one of the core drivers of creativity (e.g., Amabile, 1979, 1985).

Other work has revealed that negative affect can drive creativity under specific circumstances, such as when individuals perceive that they will be recognized and rewarded for their work and experience their mood lucidly (i.e., exhibit clarity of feelings, George & Zhou, 2002). Further, some research suggests that shifting between positive and negative moods over time may contribute to creativity (George & Zhou, 2007). Additionally, Fong's (2006) series of laboratory experiments suggests that ambivalence, defined as positive and negative feelings (e.g., Pratt & Doucet, 2000; Fong, 2006; Fong & Tiedens, 2002; Larsen, McGraw, & Cacioppo, 2001; Williams & Aaker, 2002; Rothman & Wiesenfeld, 2007), can support creativity by enabling individuals to make unusual associations.

Though much work has explored how positive and negative affect may, under various conditions, support creative outcomes, the bulk of this work has focused largely on affect as an input into the creative process, and it has largely examined affect that is unrelated to the creative work itself (cf. Amabile et al., 2005; Amabile & Kramer, 2011). Though Amabile and colleagues' (2005, 2011) studies have examined positive affect as emerging from the work itself, to my knowledge, no prior work has investigated negative or mixed emotions that are tied to the focal creative task. When emotions are elicited by tasks over which individuals typically have some degree of psychological ownership (Pierce & Jussila, 2011; Pierce, Kostova, & Dirks, 2001, 2003), such as creative work (Rouse, 2016), the mechanisms at play may be different, and negative emotions may be particularly challenging. Thus, it is unclear whether the conclusions from the majority of the studies reviewed here would hold when considering emotions that arise from creative forecasting.

Following feelings-as-information theory, individuals interpret their emotional states as information about their situation, which guides their behavior (Schwarz & Clore, 1983; Schwarz, 2002; Schwarz & Clore, 2003; Schwarz, 2011). Envisioning bringing about positive outcomes through one's own work is likely a positive emotional experience, given the link between envisioning positive future events and experiencing positive affect (Scheier & Carver, 1985, 1992; Andersson, 1996). The positive emotions that likely arise from envisioning the positive outcomes of one's work may be an important source of intrinsic motivation during the creative process itself, as positive affect is associated with a sense of making progress toward goals (Amabile & Kramer, 2011; Higgins, Shah, & Friedman, 1997; Dweck & Leggett, 1988).

However, particularly when creativity poses the risk of adverse effects on others, darker emotions may arise. Because negative affect signals that something is amiss (Schwarz, 2011), negative emotions tend to halt the current plan of action, shifting individuals into an analytical, problem-solving frame of mind (Clore & Huntsinger, 2007; Schwarz, 2011). When such negative emotions are tied to forecasting the outcomes of implementing one's own creation, they may drive the creative worker to change aspects of his or her work in hopes of remedying perceived risks and diffusing aversive emotions. To illustrate, a creative worker could forecast his or her creation having unintended adverse effects on others, which could elicit concern and render moving forward a psychologically aversive experience. This could prompt the creative worker to stop the current plan of action, think critically, and potentially change his or her behavior. Yet if the creative worker forecasts both positive *and* negative outcomes, this may elicit emotional ambivalence, or positive and negative feelings (e.g., Pratt & Doucet, 2000; Fong, 2006; Fong & Tiedens, 2002; Larsen, McGraw, & Cacioppo, 2001; Williams & Aaker, 2002; Rothman & Wiesenfeld, 2007), regarding the future of his or her work. Depending on how individuals respond to ambivalence (see Pratt & Pradies, 2011, and Ashforth, Rogers, Pratt, & Pradies, 2014, for overviews of different ambivalence responses), this may impact the creative process in different ways.

In sum, because of the relationship between affect, cognition, and behavior, and the likelihood that forecasting gives rise to a range of emotions, it is necessary to investigate how creative workers forecast the downstream effects of their creations, how this forecasting elicits emotional reactions, and how this may influence the creative process.

To summarize this literature review, existing research demonstrates that creativity involves taking the risk of developing novel ideas, and certain types of creative ideas pose the risk of adversely affecting large numbers of people when they are implemented. However, we know little about how creative workers forecast the outcomes of implementing potentially harmful ideas, the emotions that may arise, and how this may influence the creative process. Existing models of the creative process have not accounted for these dynamics (e.g., Amabile & Pratt, 2016; see Lubart, 2001, for a review), and they may therefore overlook a potentially significant influence on the creative process. Hence, I ask the following research questions: How, if at all, do creative workers forecast the outcomes of implementing their potentially harmful creations, and what emotional experiences, if any, arise? Further, how, if at all, does forecasting the outcomes of implementing potentially harmful creations influence the creative process? By uncovering patterns in how creative workers forecast the future of their work and manage the emotions that arise, this dissertation aims to extend theory on the role of prospective thought processes in creative work and on how creative workers navigate the dark emotions that may arise.

CHAPTER 3:
METHODS

Research Approach and Sampling Context

My research questions ask how creative workers forecast the outcomes of implementing their potentially harmful creations, what emotional experiences arise, and how this influences the creative process. To answer these questions, I used an inductive, qualitative approach to explore the creative work of developing new artificial intelligence (AI) technologies in real-world settings. A qualitative, inductive approach is appropriate, first of all, when there is little preexisting empirical work related to the focus of the research (Creswell, 1998; Locke, 2001; Lee, Mitchell, & Sablynski, 1999; Eby, Hurst & Butts, 2009). As discussed in Chapter 2, to my knowledge, no prior work has explored the creative process *in medias res* (as it happens) when creative workers are aware of the potential to adversely affect a large number of people through their work. An inductive, qualitative approach is also appropriate because my intent is to build theory, defined as examining previously unexplored relationships or processes (Lee, Mitchell, & Sablynski, 1999; Creswell, 1998; Locke, 2001), on the role of forecasting and associated emotions in the creative process. Finally, qualitative methods are particularly helpful for uncovering how complex sequences of events unfold in real-world settings (Creswell, 1998; Chiles, 2003; Langley, 1999). I apply an inductive, qualitative approach in a field setting because my research questions call for investigating creative work in a context in which implementation is a future reality, and creative workers therefore have a stake in its outcomes.

Qualitative, inductive research involves selecting samples and contexts that are likely to reveal the dynamics of interest (Eisenhardt, 1989; Pettigrew, 1990; Strauss & Corbin, 1998). To answer my research questions, I needed to identify a setting in which individuals were engaged in creative work that posed significant and known risks to the general public. The domain of artificial intelligence (AI) development, which involves creating digital technology that rivals or surpasses human intelligence in its capabilities, exemplifies potentially harmful creativity (e.g., Urban, 2015), defined as creativity that poses the risk of adversely affecting a large number of people through its implementation. In AI development, individuals are creating path-breaking technology by developing machine learning algorithms that "learn" from data to perform complex tasks, such as interacting with humans; detecting emotions; classifying and producing text, images, and music; and predicting natural disasters. As the examples of self-driving cars and autonomous drones illustrate, even today, AI is capable of making life-or-death decisions without human input. Yet with machine learning algorithms, the complex relationship between input (the code that is written to create the AI) and output (what the AI does) makes it difficult for individuals to predict and control what their creations will do (e.g., Lehman et al., 2018); the algorithms that power new AI tools are often difficult for even the experts developing them to understand fully.

The complexity and autonomy inherent in AI raise the potential for harmful outcomes (e.g., Friend, 2018; Tegmark, 2019). Although AI may help solve many of humanity's greatest problems, such as disease, poverty, and global warming, the implementation of AI also poses significant threats. For instance, scientist and entrepreneur Elon Musk recently said:

> I'm really quite close to the cutting edge in AI, and it scares the hell out of me. It's capable of vastly more than almost anyone knows. And the rate of improvement is exponential. …I think that's the single biggest existential crisis that we face, and the most pressing one. …Mark my words, AI is far more dangerous than nukes. (Zipkin, 2018)

The dangers of developing AI include the potential for AI tools to be weaponized, such as military applications of image recognition tools; the unintentional embedding of biases in AI systems; as well as the social and economic effects of job displacement due to AI. These dynamics, coupled with the current lack of best practices regarding the ethical development and application of AI (Murgia & Shrikanth, 2019), make AI development a particularly appropriate setting to study how forecasting downstream effects influences the creative process.

Sampling Strategy

Using the logic of purposeful sampling (Locke, 2001; Patton, 2001), which involves selecting individuals or cases that are most likely to reveal the dynamics of interest, I turned to individuals working in artificial intelligence (AI) development to answer my research questions. As is often the case with inductive, qualitative research, my research questions evolved over time, and my sampling strategy shifted from purposeful to theoretical (Strauss & Corbin, 1998; Locke, 2001) as I moved from pilot interviews into the remainder of my data collection.

In the beginning, my research question was quite broad; I wanted to understand the dynamics of the creative process when adverse downstream effects could arise. Additionally, I wanted to explore AI as an appropriate domain for a full-scale study. I began with a purposeful sample of eight informants from two AI startup organizations, and I also conducted an informational interview with a management and information technology doctoral student with relevant industry experience. Because of the exploratory nature of the pilot study, I was initially interested in understanding AI developers' experiences as well as other organization members' views of their organizations' work and the broader AI community. My pilot study sample therefore included those directly involved in machine learning algorithm development as well as several other members of the organizations (founding executives and a marketing director).

Because data collection and analysis occurred in tandem, as I refined my theoretical focus following pilot interviews, I also shifted my sampling strategy (Strauss & Corbin, 1998; Locke, 2001) based on what I learned from initial interviews. For the remainder of my data collection, to explore my research questions about forecasting the outcomes of implementing potentially harmful creations, emotional reactions, and the creative process, I sampled AI developers (specifically, machine learning algorithm developers), and rather than sampling within certain types of organizations, I sampled across organizations that varied in size and age as well as across academia and industry.

I sampled across multiple organizations for several reasons, based on what I learned in initial interviews. First, I learned that the work of AI development follows a similar process of building, testing, and refining regardless of setting. Second, the downstream risks associated with this work are consistent across settings (direct harm through dangerous applications or malfunctioning creations; indirect harm through job displacement), though certain applications of new AI technologies may of course pose particular forms of these risks that are specific to them. I also learned through pilot interviews that, because AI is such a new domain, organizations lack norms regarding how to address the potential consequences of AI. As such, I did not find that organizational settings differentially influenced how individuals addressed the risks posed by their work. Additionally, and also related to the newness of AI, I learned that AI developers are highly interconnected; they tend to rely heavily on the broader AI research community. In more established types of work, individuals may find most of the resources they need within their organization. In AI, though, individuals in any organizational setting tend to be very involved in the AI research community because this is where they find most of their AI resources, such as new machine learning techniques to try. Finally, I also learned that in AI development, the lines between industry and academia are blurred: Many industry companies house basic research, and many academics apply their techniques in for-profit companies. Because of the common processes of machine learning algorithm development, the consistent types of possible downstream effects, the lack of organizational norms regarding dealing with these effects, the interconnectedness of AI developers, and the frequent collaborations between academia and industry, I did not limit my sampling to one organization or even one type of organization.

The interconnectedness of AI developers allowed me to leverage snowball sampling, or chain referral sampling (Biernacki & Waldorf, 1981), to achieve a theoretical sample of individuals engaging in potentially harmful creative work. Snowball sampling, which is used frequently in the social sciences, involves making use of individuals' connections to others in similar contexts or roles (Biernacki & Waldorf, 1981). It has both benefits and downsides; its main downside is the likelihood of bias due to the reliance on personal and/or professional connections (Heckathorn, 1997, 2011; Goodman, 2011; Biernacki & Waldorf, 1981). However, it is an appropriate sampling technique when informants are difficult to reach and/or the subject matter is sensitive (Biernacki & Waldorf, 1981). In my case, due to the growing public discourse about the dangers of AI, AI developers tend to be wary of outsiders who might bring negative attention to their work, and my interviews included questions about possible downstream effects, which could lead to the discussion of sensitive subject matter. Hence, snowball sampling was an appropriate technique for my research.

My pilot interviews arose from "cold" emails to AI organizations, and I "snowballed" from that point onward, asking each prior informant if they could connect me with colleagues who might be willing to be interviewed (e.g., Petriglieri, Ashford, & Wrzesniewski, 2018). Throughout my data collection, I attempted to mitigate the potential bias of snowball sampling by asking informants to connect me with colleagues who had perspectives that might be different from theirs. Sometimes, informants told me outright that they thought their views might be uncommon, and they proactively connected me with individuals who thought about their AI work differently. This effort to understand as many different perspectives as possible was a constant thrust throughout my sampling, and the emergence of distinct patterns in my data provides some assurance that my data are not biased toward one particular perspective.

My final sample consists of 64 interviews with AI developers[3] and an additional five interviews with AI organization executives (including five pilot interviews with AI developers and three with executives). I stopped recruiting additional informants when I reached theoretical saturation; following Glaser and Strauss (1967), I define theoretical saturation as the point at which new data no longer contribute new information.

Data Sources

Semi-structured interviews. Semi-structured interviews were my primary data source. For pilot interviews, I used a semi-structured interview protocol (see Appendix I) that addressed the nature of AI development, focusing on informants' creative and problem-solving processes (if they were AI developers), emotional experiences, general features of the creative product, comparisons between human and artificial intelligence, and views about the future of AI. Pilot informants predicted both positive and negative outcomes arising from their work and expressed both excitement and concern, but they seemed to approach their work in different ways. This led me to the research questions that I then explored through the remainder of my data collection: How do creative workers forecast the outcomes of implementing their potentially harmful creations, what emotions arise, and how does this influence the creative process in the present?

[3] AI (machine learning), like many STEM fields, is predominately male. My sample consisted of eight female AI developers; at 12.5% of my sample of 64 developers, this is representative of the industry (Chin, 2018). Two of the executives were female.

Following the pilot interviews, and throughout subsequent data collection, I revised my interview protocol (Spradley, 1979) to address the concepts that were emerging as important to my refined theoretical focus. In the middle and later phases of data collection, this refined focus was whether and how individuals constrained themselves based on how they responded to their ambivalence about forecasting positive and negative outcomes. Because my research questions were intended to uncover individual experiences and processes, I often asked informants to recall recent experiences and to describe ongoing or recurrent processes. Additionally, because an essential aspect of the study is understanding how creative workers think about the possible outcomes of their work, several of my questions were prospective in nature – that is, they addressed informants' views about and plans for the future. The final interview protocol is presented in Appendix II.

Non-participant observation. I also attended a three-day AI conference, AI World (Boston, December, 2018), as a non-participant observer and took notes on my experiences and interactions (Miles & Huberman, 1994; Creswell, 1998). The purpose of this observational data was twofold. First, I triangulated (Jick, 1979; McGrath, 1982; Eby, Hurst, & Butts, 2009) what I learned from other data sources, such as the types of risks and benefits posed by AI development. To do so, I attended formal presentations and approached other attendees informally during breaks between sessions. When approaching attendees, I identified myself as a Ph.D. student researching AI development, and I asked questions about their experiences with managing AI developers or developing AI themselves, depending on their roles.[4] The themes that were discussed in formal presentations and the information I gleaned from informal exchanges supported the notion that AI developers were thinking about both the possible benefits (driving

[4] I did not formally interview attendees or snowball sample based on these connections, and as such, I do not include these conversations in my interview count.

scientific progress and improving people’s everyday lives) and consequences (causing direct or indirect harm through implementing complex and autonomous new tools) of implementing their work. As such, I found support for the patterns emerging from my analysis of interviews through this non-participant observation.

Second, conference attendance enhanced my knowledge of industry trends and dynamics. For instance, I learned about recent and growing applications of AI, such as in cybersecurity, financial services, healthcare, and pharmaceutical research; the widespread struggle to earn the public’s trust in new AI tools; and the increasing attention on the ethics of AI development (e.g., unbiased AI) and “AI for good” (e.g., AI that supports sustainability efforts). This knowledge gave me additional legitimacy as an interviewer and provided useful background information for my interviews (Feldman, Bell, & Berger, 2003). For example, learning about the growing discussion of AI ethics and projects deemed morally “good” enabled me to ask informants what they thought about these trends and how they related to informants’ own work, if at all.

Data Analysis

My analysis focused on my primary data source of semi-structured interviews (pilot and subsequent interviews), and I used several data reduction tools in tandem with data collection and analysis of the full, unreduced data set. First, I took notes during each interview and composed contact summaries, which are structured summaries of each interview that highlight aspects that seem important to the research questions (Miles & Huberman, 1994). Second, I drafted theoretical memos, which contained my ideas about the themes and patterns that were emerging across data (Strauss & Corbin, 1998; Creswell, 1998). Third, I kept a data table, which compiled quotations expressing similar concepts and experiences from multiple informants (Wolcott, 1994; Creswell, 1998). The data table helped me to gain a clearer sense of emerging patterns, categorize concepts and experiences in multiple ways, and examine how new data fit with existing patterns. Additionally, I used my field notes from the AI World conference as supplementary data, which enabled me to triangulate (Jick, 1979; McGrath, 1982; Eby, Hurst, & Butts, 2009) what I learned from interviews. Together, these data reduction tools served as a way of structuring my thinking and analysis, developing my ideas, and comparing themes and patterns across sources.

I coded the complete interviews using grounded theory methods (Strauss & Corbin, 1998; Locke, 2001). I conducted this analysis by iterating back and forth between the data and emerging themes and by abstracting theoretical categories from those themes (Strauss & Corbin, 1998). Data collection and analysis informed each other throughout this process; I coded interviews in batches, and I adjusted my protocol (Spradley, 1979) when needed to address emerging themes. For instance, the notion of constraints began to emerge during my analysis of initial batches of interviews, and I adjusted my protocol over time to address the presence or absence of constraints more explicitly. Below, for the sake of clarity, I describe this process in a more linear way, based on the stages described by Pratt and colleagues (2006).

Step 1: Open coding. For each batch of interviews, I began by creating provisional, "open" codes of statements that seemed relevant to my research questions regarding forecasting, emotions, and their influence on the creative process. These provisional codes often used informants' own language, reflecting back their experiences in their own terms (Locke, 2001). For example, "It shouldn't ever be allowed to make decisions on its own" was coded as *requiring human involvement,* "How do we almost make sure that it doesn't generate things we wouldn't want it to generate?" was coded as *controlling output*, "We still have to do testing, we still have to do certification, we have to do retraining, we have to look at all the implications" was coded as *testing thoroughly,* and "You should always have more people's eyes" was coded as *seeking external checks*.

As data analysis continued, concepts began to saturate, and my interview protocol evolved to ask more pointed questions about forecasted outcomes, emotions, responsibility for negative outcomes, and constraints. As such, my open codes also evolved to better reflect these emerging themes, such as the types of positive and negative outcomes individuals forecasted (e.g., "pushing research forward," "predicting weaponization"), emotions related to forecasting (e.g., "excited about practical applications," "worried about bias"), the degree to which individuals took responsibility for negative outcomes (e.g., "malicious applications seen as irrelevant," "sense of moral imperative"), and the types of constraints they did or did not impose on their work (e.g., "refusing certain projects"). These open codes encompassed concepts that were common to multiple interviews rather than specific to an individual.

Step 2: Axial coding. Next, I developed axial codes, or broader theoretical categories, that encompassed subsets of the open codes. For instance, the open codes of *requiring human involvement* and *controlling output* were categorized under the axial code of *constraining functionality*, and the open codes of *testing thoroughly* and *seeking external checks* were categorized as *constraining release process*.

As I developed these theoretical categories, I frequently returned to the data to compare and categorize individuals' perceptions and experiences. As I did so, I drafted theoretical memos and revised my data table. Over time, as I integrated new data into my data table and revised my theoretical memos, I found patterns among those who did and did not constrain their work in terms of how they responded to their ambivalence. Specifically, I found that informants differed in whether they increased or reduced their psychological distance (Lewin, 1951; Trope &

Liberman, 2003) from the negative outcomes that they forecasted, which related to whether they implemented constraints.

I sometimes refined the associated axial codes to reflect what I learned during this stage of analysis. For example, during this stage, I refined my understanding of one group of informants, who expressed largely positive emotions about their work and did not constrain themselves. I had initially categorized these individuals as unambivalent due to their excitement about the future of their work and apparent lack of concern about potential downsides. However, when I returned to the data and compared these informants to others, it became clear that they had forecasted negative outcomes like other informants, but they seemed to have resolved their ambivalence by amplifying their positive thoughts and feelings about their work. I refined the axial codes regarding their emotions to reflect this shift in my analysis: "excitement" and "embracing positive outcomes" became "amplifying positive thoughts and feelings" by "emphasizing positive outcomes."

Step 3: Aggregating theoretical dimensions and developing the process model. After developing theoretical categories, I worked toward aggregating them into broader dimensions. For instance, *constraining functionality* and *constraining release process* were aggregated into *applying self-imposed constraints*, which became associated with *concretizing negative outcomes* (envisioning negative outcomes in a detail-oriented way).

At this stage of the analytical process, I also began developing a process model that encompasses how forecasting positive and negative outcomes elicits ambivalence and how individuals respond to this ambivalence in different ways, which relate to the role of self-imposed constraints in their work. The purpose of the process model is to extrapolate from the data the theoretical dynamics that have emerged from one's analysis – to show why the data matter in theoretical terms – rather than to reiterate a descriptive account of the findings (Langley, 1999; Chiles, 2003). As such, I pushed myself to move beyond the specifics of my data and to think carefully about how concepts seemed to fit together. For example, I theorized that the process of concretizing negative outcomes arises from commitment (Brickman et al., 1987; Pratt & Rosa, 2003; Pratt & Pradies, 2011) to one's work and motivates the application of self-imposed constraints, and I aggregated these concepts into the novel "redistribution" response to ambivalence.

CHAPTER 4:
FINDINGS

Overview of Findings

Creative ideas are inherently novel (Amabile, 1983, 1988; Oldham & Cummings, 1996; Woodman, Sawyer, & Griffin, 1993; Stein, 1974), and creativity therefore carries the risk of unpredictable outcomes. Some domains, such as artificial intelligence (AI) development, are marked by a particularly wide scope of impact and a heightened potential for harming the general public through the implementation of new creations. Under such circumstances, individuals may grapple with the negative downstream effects that could arise from implementing their work. As I have argued in the previous chapters, forecasting the outcomes of implementing creative ideas likely elicits emotional reactions, which may influence the creative process.

Through an analysis of interviews with 64 AI developers and supplemental data (field notes and five interviews with AI organization executives), I find that individuals forecast both positive and negative outcomes arising from implementing their potentially harmful creations. This is associated with ambivalence, defined as "simultaneously positive and negative orientations toward an object" (Ashforth, Rogers, Pratt, & Pradies, 2014: 1454; Pratt & Pradies, 2011; e.g., Pratt & Rosa, 2003), and how individuals respond to this ambivalence is related to whether they choose to constrain their creative work as it unfolds. Because my findings center on the role of self-imposed (rather than extrinsic) constraints, I adapt Cromwell and colleagues' (2018) definition of constraints to reflect constraints that creative workers impose on their own work: I define the term *self-imposed constraints* broadly, as any limit or boundary that individuals impose on their creative process or its outcomes, at any stage of the creative process (i.e., from choosing the types of projects that they will pursue to determining the process that they will use to test, refine, and implement their ideas).

A Common Starting Point: "Noisy" Forecasting, Positive and Negative Outcomes, and Ambivalence

"Noisy" forecasting. Regardless of the role of constraints in their creative process, all informants in my sample explained that the nature of their AI development work made it difficult to predict what their creations would do, let alone the downstream effects that could emerge from implementing them. One informant, Lance, explained:

> It's one of those things where you don't know what you don't know. …And you hear similar stories with people who are working with AI in all kinds of industries. Like, what happens when we just turn it on and let it do something for a long time? You just get surprised at the output.

Richard reasoned that the difficulty of predicting what AIs would do was related to the nebulous relationship between input (the code that is written to create the AI) and output (what the AI does) that marks machine learning algorithms: "…The relationship between what you put in and what you get out is so complicated, that you don't really have a rigorous understanding of what's going on necessarily."

This uncertainty about how AI creations may behave in the moment is related to uncertainty about more distal outcomes. Lance likened the domain of AI development to driving at night:

> …[Developing AI is] sort of like driving home at night, like you can only see 30 feet in front of you where your headlights are. But nevertheless you can successfully drive all the way home only by seeing 30 feet in front of yourself. And that's kind of like what this feels like a little bit – like you're driving along, and at any point in time, we can only see two-ish years in the future, we think. But as we march along… that space of our headlights that we can see [goes] a little bit further. That's what it feels like to me.

The creative process of AI development is infused with a sense of not knowing what one's work will become or how it will be used. To add nuance to the metaphor, what exactly "home" is remains in question. Informants also made observations about the broader field of AI development that cued them to the potential for sudden changes. Zack commented on the relatively frequent emergence of new paradigms in AI:

> Once in a while, and within the next year or two years, someone will come up with a new idea, which will just completely change again. Just like this. The point I'm sure of, the technology is yet to be completely explored. …Maybe there will be something completely different from deep learning in the next year, possibly, which will trigger a paradigm shift, which we have not foreseen before.

As a second example, Doug commented on his impression that AI development is in a period of acceleration; he reflected on learning that surprises could emerge in the field at any moment:

> What I see is that we are facing an acceleration right now. …At first I was surprised by many things, but now my state of mind is basically, okay, I'm *so* surprised by what's going on, what will be the next thing? How big will it be? How exaggerated or maybe how extreme will it be? I'm kind of making jokes on that somehow. It's kind of getting completely nuts. …And I have some people close to me who are kind of taking bets – how big will it be? AI is really surprising right now. …It will change many things. But I would say in the five last years, it's really kind of moved up in a speed that I've never seen, I was not expecting.

Hence, the constant potential for surprises in one's own work as well as major field-level changes create a sense of uncertainty about how one's work will develop over time and about the future of AI more broadly.

Positive and negative outcomes. This uncertainty about the future made it difficult for informants to trust that their work would have only desirable effects once implemented. Rather, because of the risks that are inherent in developing complex and autonomous AI tools, informants commented on the potential for new AI technologies to have both positive and negative effects on the general public. Informants sometimes referenced the development of nuclear energy as an analogy. Nelson said, “There definitely can be many scary ways to use a knife, and it will be misused. And I think the potential here (with AI) is as bad as like nuclear weapons and so on. It’s definitely another magnitude of problems that it can create.” Because of this heightened risk of uncontrollable and deleterious consequences, informants forecasted both positive and negative outcomes arising from implementing their work.

Informants discussed, first of all, the potential for their work to bring about the positive outcomes of driving scientific progress as well as improving people’s everyday lives. As an example of driving scientific progress, Jenny explained her enthusiasm about contributing to science through her work: “[I’m involved in] figuring out algorithms and figuring out the science part. I see less of a concern, and more excitement there actually, because it opens new ground for us to dig.” She was particularly interested in contributing to the invention of human-level artificial general intelligence (AGI; AI that performs at or beyond a human level on multiple tasks): “I wish I could live long enough to see [human-level AGI]. I was born too early. It’s an opportunity. That’s what keeps you going. It’s great to be part of something as big as that, to be able to be part of the creation or the process that eventually ends up there.” As an example of improving people’s everyday lives, Jack described his belief that AI would enable people to pursue more worthwhile endeavors by replacing labor with machines: “We’ve automated all this stuff, and people just don’t need to do that crap anymore. So I think AI is maybe going to be one of the most powerful examples of that. It’s going to be different than the internet or the car. This is going to get to the point where it can do 90% of what an average person can do, and that’s going to free up a lot of hopefully productivity for people.”

Yet informants also discussed adversely affecting large numbers of people through their work because new AIs could be applied in dangerous ways, unintentionally malfunction (e.g., if bias becomes embedded in the AI), or displace segments of the workforce due to automation. Hence, they forecasted precipitating forms of both direct harm (dangerous applications, malfunctioning creations) and indirect harm (job displacement). By direct harm, I refer to circumstances in which the AI tool itself is used in dangerous ways (e.g., weaponized AI – dangerous application) or causes harm autonomously (e.g., AI that behaves in undesired ways, such as due to bias – malfunctioning creation). By indirect harm, I mean that the harm is not caused directly by the AI, but by its usefulness as a tool, which can lead to replacing human workers in its domain.

As an example of forecasting direct harm through dangerous applications, Jack predicted malicious applications of generative AIs, a type of AI that he was involving in developing. He forecasted:

> I think there's an amazingly horrendous thing that's going to happen. …There is going to be, if not this very second today, if not today, in the next ten years, easy – probably more like five – I guarantee my life, in 10 years, the generative AI is going to be so good, that the fakes will be impossible to determine by humans and maybe will start to fool best forensics labs in the world. What's going to happen is that whoever owns a supercomputer, and whoever has a team of 20 ML experts, can now produce fakes of the most famous celebrities or presidents doing whatever they want, and it's impossible to prove that that's not an original. So what does this do? This destabilizes news, politics right? …Our society is completely unprepared to deal with a technology this powerful. It's going to destroy so many things. It all comes down to, is there going to be an evil criminal gang that decides they can use this? I guarantee somewhere in the world. And they're going to be amazingly powerful.

To illustrate forecasting direct harm through releasing a malfunctioning creation, Christopher discussed the dangers of releasing an AI that had been inadequately trained:

> I remember thinking about, like, "Oh, self-driving cars, that could be huge thing," where like all the training data is all of like normal looking bipeds, and then you have training example of maybe a kid on scooter, or someone in a wheelchair, and that just isn't part of the majority class… I wonder about all those things. I wonder about those risks.

Informants also explained how their work could displace segments of the workforce by automating processes currently done by humans. As an example of forecasting indirect harm through job displacement, Matthew predicted:

> I don't buy this thing at all that everybody will just find something better to do. That's not true. That's not true at all. I'm not a business person, but I think there's plenty of people lost their jobs at Ford Motor Company Plant in the 80s who were never really gainfully reemployed. They didn't become factory automation engineers, they just became unemployed people. For sure there are people that are gonna be displaced. I think that it will serve as a tool of increasing economic inequality. I think for sure.

Informants described job displacement as potentially precipitating an upheaval of the economy.

Ambivalence. As informants expressed their uncertainty about the future and forecasted both positive and negative outcomes, they expressed ambivalence, defined as "simultaneously positive and negative orientations toward an object" (Ashforth, Rogers, Pratt, & Pradies, 2014: 1454; Pratt & Pradies, 2011; e.g., Pratt & Rosa, 2003). "Orientations" involve both thoughts (cognitive ambivalence) and feelings (emotional ambivalence), which often arise in tandem (Sincoff, 1990; Ashforth et al., 2014). In the present case, the "object" is informants' AI creations, including their possible outcomes. As an example of the cognitive ambivalence that arises from forecasting both positive and negative outcomes, Ryan explained:

> I have to I guess acknowledge that what I'm doing could be harmful, but then again I'm not sure… It also could be helpful in a lot of ways. That's what I think drives progress in AI is to a large extent, is all of the benefits that it brings to the table. So I think it's like a lot of technologies, where there's potential harm for many things that we introduce. We're seeing all kinds of unintended consequences all the time.

Hence, he expressed the possibilities of precipitating harm as well as making a positive impact through his work. He reasoned that any type of work may have unintended consequences and that it was important to drive progress. As an example of both cognitive and emotional ambivalence, Jordan expressed, "It's difficult to guess what will happen. It definitely resonates with me that you can only see, you know, see that far into the future. Which, I mean, on the other hand, that does not bring me fear. It brings me an equal amount of fear and excitement, right?" Jordan's emotional ambivalence seemed to arise from considering the possibility of a wide range of outcomes. Of course, cognitions and emotions influence one another (Sincoff, 1990; Ashforth et al., 2014), and it is impossible to separate cognitive and emotional ambivalence in my data. As such, throughout my dissertation, I use the term "ambivalence" to refer to positive and negative thoughts and feelings.

Ambivalence Responses and the Role of Self-Imposed Constraints

As informants confronted the uncertainty inherent in their work, they reckoned with the potential for both positive and negative effects to emerge, and they experienced "mixed feelings" as they forecasted "mixed outcomes." How informants responded to this ambivalence had important implications for their creative process in terms of the role of self-imposed constraints in their work. Below, I present quotations from informants to support the theoretical themes and processes that I have induced. Additional representative quotations that illustrate these concepts – as well as the concepts discussed in the prior section – are presented in Appendix III.

1. Unconstrained Creativity Path

One group of informants moved forward in their work without imposing constraints. These informants amplified their positive orientation (thoughts and feelings) toward their work and its future, rendered negative outcomes irrelevant to their day-to-day work, and saw constraints as unnecessary or as impediments.

Amplification of positive orientation. This group of informants amplified their positive thoughts and feelings toward their work by creating psychological distance (Lewin, 1951; Trope

& Liberman, 2003) from the negative outcomes that they forecasted. They did so by anchoring on the present moment and/or by emphasizing the positive outcomes that they forecasted.

Anchoring on the present moment entails focusing on aspects of daily work, such that negative downstream effects are rendered irrelevant, whether by focusing on the technical details of work and the perceived boundaries of one's role and/or on the limitations of the current technology. To begin, some informants described their absorption in the technical details of their work and their belief that responding to potential negative outcomes was not their responsibility.[5] Bob expressed zeroing in on current technical challenges rather than weighing distal outcomes, and he seemed to dismiss the connection between his work and adverse downstream effects:

> For researchers like me, we do not care too much about what the future might bring. Instead, we are looking into the high-impact or the burning questions we are having today and trying to answer it. Where it will lead us, we actually do not care too much. As soon as we know the problem we are attacking is meaningful, it's important. Maybe we will go nowhere, maybe we actually will have Skynet. But for researchers like me, we do not… personally, I do not worry too much.

He elaborated that although he was aware of the potential for negative downstream effects, he did not see constraints as part of his role: "There are a whole lot of other people who are worrying about the regulations, the security, about fairness, but they are in different and separate research areas. I do agree that these topics should be taken care of, but it's probably by someone else."

Informants also anchored on the present moment by reasoning that the limitations of the current technology made it unlikely that their work would precipitate negative effects within a relevant timeframe. Several informants, such as Sean, commented that current AI creations were not as sophisticated as people thought:

[5] One informant speculated that focusing on improving his technical skills would allow him to better prepare for the threats he expected, such as job displacement and malicious applications of new technologies. However, no other informants described thinking this way.

> Ultimately what the systems are doing, at least when they're learning, is they're optimizing the functions using gradient descent usually. So gradient descent has been around since Newton. And so it's not all that spectacular. It's basically multivariable calculus. And then we're applying that with more data than we used to have, and we can do that because we have slightly better computers than we used to have. And so combine that with humans, who tend to take the intentional stance about everything, and it's very easy to think that AI is super awesome, but actually it's very rudimentary multivariable calculus.

He explained that, because he believed that AI was so rudimentary, he was not afraid of the future of AI:

> I think it's important to just realize how insignificant everything is and how far we still have to go. I guess that's also the answer for dealing with the fear aspect. I think people grossly misunderstand the capabilities of modern AI systems and really think that strong AI is around the corner, which I definitely don't think it is. So machine learning is just fitting a bunch of parameters to a maximum likelihood estimation. So you can learn some statistical properties of the world, but doing more sophisticated things is pretty much still beyond this I think. So we still have a long way to go.

Informants taking this stance also expressed that, because they did not expect major changes to take place soon, they did not feel guilty about playing a part in the negative outcomes that they forecasted:

> I don't feel that guilty that much right now. …At this point in the evolution of AI, the displacement is small, and it's going to stay small for a long time. For sure, the things that are going to be displaced right now, yeah, it's going to be more rote things, things that require a low degree of creative input or whatever. But I think this is gonna increase, right? It's not gonna go away. It'll be a tiny percentage every year forever. It will have a wide impact. It's that old thing about boiling the frog – the frog doesn't notice that he's in boiling water because it's only raising a tenth of a degree every minute. So it isn't gonna be a sharp spike in displacement, it's just gonna happen slowly over time. (Matthew)

Informants also created psychological distance from negative outcomes by emphasizing the relative importance of the positive outcomes of driving scientific progress and/or improving people's everyday lives. As an example of emphasizing the positive outcome of driving scientific progress, Dennis explained his view that, although his creations could be used in dangerous ways, contributing to science through his work was more important:

> Yeah, my work is linked indirectly with this [threat of weaponized robots]. …My work is one more brick in the building of adaptive robots and autonomous robots. If you want to have efficient robotic soldiers on the battlefield, you need the kinds of things I do. Of course. But if we stop to work in science because people will use it in not a good way, we stop to do science. For me it is not like… It is not a concern for me.

As an example of emphasizing the positive outcome of improving people's everyday lives, Jenny expressed her optimism about replacing human labor with AI:

> I'm very optimistic about this job thing that people talk about. Because I actually have a vision in the future that people don't have to work, because the wealth will be grown by automation and all the advanced technologies. So not having a job is not a bad thing. And we can eradicate poverty, we can eradicate disease, we can give ourselves more purpose, we can make ourselves more robust, we can go to other worlds, live on the moon or Mars or something. And all that I think has to do with us building these technologies that make it possible. …You get to this superhuman that doesn't have diseases and is not poor and can concentrate on positive things, construction, be creative.

Informants who emphasized positive outcomes often reasoned that any negative effects that might arise from their work, such as the weaponization of their creations or job displacement, would be outweighed by attaining the positive outcomes that they forecasted.

By creating psychological distance from negative outcomes, these informants amplified their positive orientation toward their work, therein resolving their ambivalence. To illustrate, Andrea expressed her optimistic view of AI, the importance of her AI research work, and her hopes that AI would become ubiquitous:

> I think I'm very optimistic about the future of AI. I think at this point it's all about research. So you cannot find any example of, like, cars or something for autonomous driving – I think we're quite far from that. But maybe in a few years, the research will be still ahead, but industry will catch up. …So, yes, I am very optimistic. …What I hope to see happen in my lifetime is AI becoming, being used, so you could find it all around you.

Hence, despite the fact that they forecasted negative outcomes arising from implementing their creations, these informants were ultimately largely optimistic about the future of their work and AI in general.

Unconstrained creativity. Informants who amplified their positive orientation toward their creations viewed constraints as unnecessary or even as impediments to their work. For instance, Andrea voiced her belief that constraints would be counterproductive, reasoning that it is necessary to take risks in order to progress:

> There was this example from Facebook, I think. They tried to get two neural networks, two AIs to talk to each other, and at some point, they developed their own language, and the researchers couldn't understand them anymore, so they just shut them down. I mean, so what? [Laughing] Let them talk! …There are risks of course, but if you don't try it, the research will never progress.

Nick voiced a similar perspective, fearing that constraints would halt scientific progress: "You cannot control everything, and so you don't know what everybody will do with AI. And sometimes this will also have the negative effect of people trying to pose resistance to AI and slowing down the research process." Hence, some informants saw constraints as an impediment to their ability to engage in their work and achieve the positive outcomes that they forecasted.

Engagement in work: These informants enjoyed their work and seemed immersed in the technical challenges before them, and as such, they may be described as engaged in their work – that is, exhibiting involvement in, satisfaction with, and enthusiasm for their work (Harter, Schmidt, & Hayes, 2002). For instance, Sean said, "I think AI is very interesting, so if you take away all the hype, it's basically the most interesting problem I think you can work on because it's about the most fundamental question you can ask as a scientist, almost, which is what makes humans human." Jacob described finding enjoyment and awe in his AI development work:

> I was experimenting with this music app that I was building, and all of a sudden it started feeding back on itself, and then the feedback started feeding back from

> itself, and all of a sudden it almost sounded like something was speaking to me. It was terrifying and amazing at the same time, and it was such a simple algorithm that was running. I have never felt that way doing traditional software development, but I'm starting to get inklings of that again now, where it surprises you in ways you don't expect. …It's the kind of thing where the solution that it gets is so unlike yours that it just, it's mind-blowing to me.

To summarize the "unconstrained creativity" path, these informants seemed able to engage in their work because they had created psychological distance from negative outcomes and amplified their positive thoughts and feelings. As such, they moved forward without imposing constraints on their work.

2. Heedful Creativity with Self-Imposed Constraints Path

Though informants discussed in the prior section did not constrain their work, the majority of my sample (46 of 64) responded to their ambivalence in a different way. These informants were committed to their work (Brickman et al., 1987; Pratt & Rosa, 2003; Pratt & Pradies, 2011) in that they accepted that both positive and negative outcomes would likely occur. Though they were simultaneously excited about positive outcomes, these informants tended to concretize the negative outcomes that they forecasted – to envision them in a detail-oriented way. Concretizing seemed to give rise to negative emotions like worry and concern, temporarily amplifying informants' negative orientation toward their work. However, these informants took responsibility for negative outcomes by imposing constraints on their work, which enabled them to "redistribute" their ambivalence and strengthen their positive orientation toward their creations. Because these informants thought critically about the negative outcomes that they forecasted and responded by approaching their work with greater caution, I refer to this path as entailing "heedful" (Ryle, 1949; Weick & Roberts, 1993) creativity.

Commitment to work. These informants accepted that both positive and negative outcomes would emerge from implementing their creations. To illustrate, Ray expressed his acceptance that positive and negative outcomes would arise from creating a “powerful tool” like AI: “When you have a very powerful tool, there’s going to be good and bad consequences. There are good and bad people using it. We can come up with something that recognizes who a person is, what they are intending to do – that can be used for good or bad.” As another example, Gary reflected on his belief that AIs would solve hard problems – and likely be applied in unethical ways as well as displace human jobs:

> I think misuse of AI and a lot of applications which are maybe ethically arguable are going to come out. Because these models are getting better and better at human intelligence tasks, like recognizing faces. A very useful application for a military organization would be to recognize the face of someone they want to kill and then kill them. That’s very good for them. …So I guess the way I see it is optimistic, in that we’re going to keep solving these hard-to-solve problems that no one thinks a machine can do, and we’re going to do it more and more accurately. And negatively because I think that corporations are going to be using these models in ways that are going to net them revenue. …So I guess that’s how I see AI going, is replacing a lot of [jobs] – or even introducing new ways for people to automate at a large scale these human intelligence problems. And that to be used by corporations and the government for both positive and negative.

Frank echoed this acceptance of positive and negative outcomes arising from his work and noted his plan to continue to develop AI tools:

> I really do, I believe there’s some both really exciting and scary possibilities when you combine machine learning and blockchains, if you have like blockchains that have intelligence baked in about how to change themselves. …It’s definitely something I’ll keep using.

I describe these informants as being committed to their work, drawing on Brickman and colleagues’ (1987) conceptualization of commitment in terms of the binding of positive and negative orientations toward an object through choice. In the present context, AI developers bound their positive and negative orientations toward their creations by accepting the positive and negative outcomes that they forecasted. This commitment seemed to manifest in two cognitive and emotional processes: excitement about positive outcomes (tied to acceptance of the positive) and concretizing negative outcomes (tied to acceptance of the negative). Hence, though I have introduced the notion of commitment in the present section, informants' commitment to their work is implicit in processes described below.

Excitement about positive outcomes. Informants expressed enthusiasm for the positive outcomes that they forecasted and were dedicated to their work as a way of achieving them. For example, Marie explained her view that AI would help improve people's lives in multiple ways:

> AI can really help us to better our understanding of the world. It can help us to spend more time on interesting things and can help us to become more intelligent. So for me, AI can really give some positive, very positive applications for us. And also, there are AI, for example, to help poor people improve their life conditions. We can also say that AI can help us to become – to don't have any problem (sic), for everyone, it could be possible. So this kind of motivation, still it is quite important for me.

Joe echoed this sentiment: "I can see a future where AI will help people in many things in life, and will replace many jobs, and maybe will improve many things in society." Steven expressed his "love" for his job as a way of solving complex problems: "That's why you love this job, are the unexpected outcomes you end up having, simple ideas that actually make it to algorithms that solve complex problems. They solve complex problems, but with surprisingly simple solutions. And this is what makes things exciting."

Concretizing negative outcomes. Just as these informants were excited about realizing the positive outcomes that they forecasted, though, their commitment also manifested in their acknowledgment of the negative sides of their work. Rather than increasing psychological distance from potential negative outcomes, they tended to envision these outcomes in a detail-oriented way, which I refer to as *concretizing*. Through concretizing negative outcomes, informants seemed to (temporarily) amplify their negative thoughts and feelings about their work and develop a sense that they needed to address the negative outcomes.

Concretizing outcomes appears to have roots in the process of assessing how a nascent idea could develop. To illustrate, Thomas explained his tendency to envision what his creations could become:

> Maybe the deeper thing here is a little bit about potential. Research-grade stuff is pretty ugly, it's aesthetically unappealing. In my particular case, it's a little 2D world, and it's almost like stick figures, and that kind of… In my mind, I'm seeing the sort of "dot dot dot," maybe reskinning this, looking like 2D blob characters, or having a cartoon sort of version of it. …There's some core idea, and it has a certain amount of potential. And you're always trying to assess, basically, what's the maximum potential of this thing?

This tendency to see the potential of an idea includes envisioning the potential downstream effects that could emerge from implementing it. Ray explained how future scenarios once deemed purely science fiction now seemed more realistic, which made the risks of implementing new inventions "real" as well – and created a need to address them:

> …You see that, "Wow, this might actually happen." But your perception of it becomes more broad. You start seeing that, well, there are things that we now have to deal with. We didn't think of biases before, now we have to think about it. We didn't think of safeguarding and making systems robust, we have to think about it next. We have to think of how we actually combine human brain processes with mechanical processes… So there's a broadening of that. …Now it's like, "Yeah, there's more concrete, cool things – there's also things that we have to deal with to make sure they don't happen." They become more concrete. They become more clear and more real.

As an example of concretizing, Zack described his fears about how his creations could be used by those with malicious intents:

> For what I do, with the ethics is also there. …It's a very disruptive technology, but, I mean, mastering the atom bomb was one. …Yes, I guess, I now can make some tools which would be very useful for dictators. And, okay, I was kidding about Trump, but if you look at what's happening in China, I mean, in China, there is a really a big push to apply these techniques to everybody, so I'm not sure that's a good thing. But if I were to be afraid for the future, okay, AI is part of it. …There is a big problem with ethics. …AI, in the hands of a dictator, is… the Second World War would have been very different if there was this technology.

As he envisioned how possible adverse effects might play out, such as by emboldening dictators, Zack seemed to see the connection between negative outcomes and his own work.

When potential outcomes become more concrete, psychological distance (Lewin, 1951; Trope & Liberman, 2003) is reduced; concretizing seemed to motivate informants to take action to reduce the likelihood of negative outcomes. Christian voiced this feeling of responsibility for the future of his creation: “It’s feeling responsible for what I was doing. …I think it really stems from my strong belief that when you develop a system, a technology, whatever it is, you carry at least part of the responsibility of what this technology is going to become.” Christian considered partaking in risk mitigation to be a “moral imperative.” He said, “If you are not part of this, in my opinion, this is where you would be responsible of those outcomes, because you didn’t want to look. It’s a kind of moral imperative, if I may say, to really be aware of the possible consequences of the technology that you helped develop.” Adam reflected on his newfound sense of responsibility for the potential downstream effects of his work:

> The more I think about it, the more I think there is some responsibility to make sure that you’ve at least thought a little bit about that. …Like otherwise where does the moral responsibility lie? …Maybe we’d all be better served if we thought a little bit about the downstream effects, the system we’re embedded in, and what we’re contributing.

As informants concretized the negative outcomes that they forecasted and developed a sense of urgency to take responsibility for them, their negative thoughts and feelings about their creations seemed to temporarily amplify. Adam described his “deep concern” about the future and his desire to answer the question, “How do we improve our situation without too much risk of hitting a place where the outcome is catastrophic?” Though he was enthusiastic about realizing positive outcomes through his work, he was deeply concerned about potential adverse effects. Adam expressed that he felt primarily pessimistic when thinking about the future of his work: “I’m feeling a little pessimistic about [the future], I guess [laughing].” He specified that his fears focused on the risks associated with creating an intelligence that could surpass humans: “…If we did create something truly intelligent, then there is some risk of displacing what actually enables our own evolutionary success. So we were flexible and smart, but then there’s something smarter than us, and that could be risky.” Similarly, Kevin expressed his fears about others’ implicit trust in AI, his fear about the future, and his resulting cautious approach to his work: “I don’t think I’m very excited. We’ve put a lot of trust in these algorithms, and as I think you’ve gathered, we don’t fully understand how they work, why they work. …So I think I’m less excited and more cautious in many cases.”

Heedful creativity with self-imposed constraints. As informants considered the negative outcomes of their work in a detail-oriented way, the amplification of negative thoughts and feelings appeared to compel them to take action. As an example, Barry expressed his concerns about his team’s technology being used in potentially harmful ways and his desire to put a stop to the current plan for his work:

> I got mad at people because they put up this blog post about how these pre-crime prevention programs were the new cool thing to use deep learning for, and it was like, “Dude, why are we promoting this? This is, like, an insane thing to be doing.” We should really be taking a step back and asking, “Why should people be using the technology this way?” Because in the long run, people are probably going to say, “It shouldn’t be used this way.” So let’s just take five minutes to think. Just cause it’s new and unregulated and doing crazy new things doesn’t make it all good. My new kick that I’m on is like, where are they using this stuff that they really shouldn’t be?

Due to their concern regarding the negative downstream effects that they concretized, these informants used self-imposed constraints to take responsibility for negative outcomes, which seemed to enable them to move forward in their work with significantly less concern.

Self-imposed constraints take a number of forms, which are generally tied to forecasting particular types of negative outcomes, and they occur at particular stages of the creative process. Below, I describe each type of constraint within three broader groups that reflect the purpose of each type of constraint: eliciting a sense of positive moral valence, control over the creation, or trust in its quality. By implementing constraints that served these purposes, informants seemed to develop a sense that they had placed necessary boundaries around their creative process and/or its outcomes and that they could explore openly within these boundaries. I describe these self-imposed constraints independently of one another, but they are not mutually exclusive; informants often constrained their work in multiple ways. Table 1 shows how I have categorized these self-imposed constraints in terms of the purpose they seem to serve.

Table 1: Types of Self-Imposed Constraints Categorized by Purpose

Positive Moral Valence	Constraining Project Choice		
Control	Constraining Functionality	Requiring Full Understanding	
Quality	Constraining Input	Constraining Own Reactions	Constraining Release Process

Positive moral valence: One method of taking responsibility for forecasted negative outcomes entails refusing to pursue projects for which one forecasts causing direct harm through dangerous applications. *Constraining project choice* enabled informants to develop a sense that their creations had a positive moral valence – that they were inherently good or safe. This constraint takes place at the beginning of the creative process, when individuals are identifying the problem or task that they will address (Amabile, 1983, 1988; see also Lubart, 2001).

Informants described choosing not to take on projects for which they forecasted dangerous applications, such as in the military. To illustrate, Thomas recounted:

> There was a … Defense project, and all they wanted to know was to count the number of cars that were in a city using aerial photography. And it's just like, I

> personally just didn't want to be anywhere near that. …It doesn't take too many hops of the imagination to imagine why you'd need to count cars from aerial photography. …I was like, "I can't be on this project, I can't do stuff – I don't want to be involved in aerial photography." So that was a decision basically on the set of problems that I'm interested in working on.

Similarly, Henry described forecasting dangerous applications for a particular technology and consequently deciding not to pursue that line of work:

> One recent thing, I said I will not work on that because I think it's too dangerous. There was a car, and that car was sponsored by Defense, and they say things like, "We want to get real-time video processing with AI capacity for embedded systems." And I said, "No." No, no, no, I will never touch that. Because I see that this is a direct image that will support drones, like smart drones, because you know if it's able to process video in real-time, then you could just put this on drones and be able to make object/face recognition with this. This is one key technology that will lead to dangerous things like drones that will… So I just answered, "No." Because someone asked me, "Would you be interested to work on that?" And I just said, "No, no, no, no, no." I will be really far away from that. I will never touch these things.

Constraints on project choice sometimes drove informants to focus on a relatively narrow subset of challenging problems that had a positive moral valence. Adam, mentioned earlier, described his "ongoing struggle" with wondering whether his work would make a positive contribution to the world or if it would be used in a malicious way. He decided to constrain his choice of projects to focus explicitly on developing machine learning algorithms that would be used as AI safety tools:

> These days I've switched my focus of research to a subfield called AI safety. And some of the concerns with AI safety are very practical, about making sure that algorithms do what we expect them to do. …You probably don't want, I don't know, a sidewalk package delivery robot to probably be surprising you and running onto the road or something, because that would be bad. …And there's a grander sort of facet of AI safety, which is – if you're familiar with Nick Bostrom, a philosopher who's quite concerned about the existential risk of AI for example. …I think it's worth thinking about. I'm also getting more into the AI ethics space, the ethics of AI and making sure that the AI behaves ethically. And making sure that we understand what that means. …For now, [focusing on AI safety] is sort of a nice resolution to my ongoing existential crisis. At least I feel like I'm pushing on the positive valence sort of outcomes. …I feel like I have

> more confidence that I at least understand – that I'm trying to do something useful.

Hence, choosing to pursue projects that are expected to benefit and protect humanity provides a powerful antidote to the fear that can arise when confronting the possibility of harmful outcomes for certain types of projects. By choosing projects that they believed were inherently good or safe – and by declining projects that they believed would be used in dangerous ways – informants developed a sense that their creations would be beneficial for the general public.

Control: Informants took responsibility for various concerns about implementing their work by imposing constraints that provided a sense of control over their creations. Control constraints arose from forecasting direct harm through dangerous applications and/or malfunctioning creations as well as indirect harm through job displacement. As Steven explained, "We are moving on the right direction by increasing the awareness that the more AI is capable of doing things, the more we need to be cautious about controlling what it does." Two types of constraints engender a sense of control over the creation: *constraining functionality* and *requiring full understanding*. Constraining functionality and requiring a full understanding occur in the middle and final stages of the creative process, respectively, as individuals grapple with the desire to wield authority over creations that are capable of acting autonomously.

Constraining functionality refers to restricting the capabilities of the creation in some way, such as requiring the involvement of a human decision maker, to prevent the AI from acting autonomously in dangerous ways or to prevent job displacement. This constraint takes place during the idea elaboration stage, when a particular idea is being further developed and refined based on its forecasted implementation (Csikszentmihalyi, 1997; Mainemelis, 2010). As an example, Barry reflected on the "scary" thought of fully autonomous AI and the importance of human involvement in the use of his technology: "I mean, [fully autonomous AI] is a scary thought. It's helpful to be shown what the thing thinks that you should do, but in my opinion, it shouldn't ever be allowed to make decisions on its own without some level of human oversight." Other informants echoed this desire to place limits on the functionality of their creations based on concerns for how humanity might be affected if they were to pursue total automation. Alan spoke of his personal evolution toward combating the possibility that his creation would be used to replace humans:

> An expectation that became increasingly important as time went on was not just that it wasn't yet capable of replacing humans. It moved from that to me wanting to *express* that I didn't want it to replace humans. And then it's moved from that lately to I am actively *working against* the ability that it could ever replace humans. So now I consider part of the goals of my research are to actually make my work resistant to that function. So even if someone read my papers and wanted to build a similar system, it would not do what they want it to.

He described working toward this goal by ensuring this his creation would "maintain a dependence on human beings," such as by requiring human input at different stages. He also described his goal of intentionally making "clumsy assistants," such as by restricting the speed at which they produced output:

> My goal is now to have it take a month to make a game. Like, I want it to really take forever. So instead of trying to solve the problem of how does [Game Design AI] generate art, it's going to commission an artist, and it's going to send them a spec, and it's going to hopefully negotiate what kind of art it wants. …And it's not just that it's dependent on humans; it has to wait for humans. So it can't, like, make 300 games in a day, because it has to wait, like, a week for the artist to email it back. So I'm kind of engineering slowness into the system, and I like that.

Hence, Alan addressed his fears regarding job displacement by proactively incorporating human involvement and restricting the speed at which his AI would produce output.

Informants also described necessitating developing a complete understanding of their creations so that the AI could be reverse-engineered – and thereby stopped from proceeding further – should harmful outcomes begin to arise. Requiring a full understanding occurs just after the outcome assessment stage – that is, once one has decided to move forward with implementing a particular idea (Amabile, 1983, 1988; Amabile & Pratt, 2016), and it entails preventing oneself from implementing the idea before a full understanding is achieved. Steven exemplifies this strategy for creating a sense of control over the creation:

> We need to … make sure we understand what AI is, then be transparent with regards to the algorithmic processes of AI, so we can actually reverse-engineer if necessary, right? Otherwise, if the algorithmic process is sitting in a black box that we don't understand, then it's not necessarily a good idea, in the long run. That's what I'm saying. We need transparency, as much as possible. Then, we would be able to situate AI where it belongs, or where we want it to belong, by controlling the procedures. The problem starts when AI hides [laughs] – starts hiding information from us, and transparency is lost, and I think this is a problem.

As Steven struggled with uncertainty about the outcomes that could arise, such as AIs that "hide" information from humans, he resolved that the best way to manage this was through working toward developing a complete understanding of the AI, which gave him a reassuring sense of control.

Quality: When forecasting the negative effects that could result from releasing a malfunctioning creation, informants also implemented constraints that allowed them to trust in the quality of their creations. Quality constraints include *constraining input*, *constraining own reactions*, and *constraining release process*. Similar to control constraints, these occur in the middle and late stages of the creative process. By taking steps to ensure that their creations would be robust, high-quality tools, informants developed trust that their AIs would perform well and ultimately help – rather than harm – the general public.

Constraining input refers to determining that the data that go into the creation – and directly affect its output – are deemed safe. This involves taking extra steps during data processing to ensure that biased or otherwise unsafe data are identified and excluded to the extent possible. How *safe* is defined varies depending on the nature of the AI, but most commonly, constraining input refers to reducing bias or using non-sensitive data. This may occur during the idea validation (evaluation) stage (checking idea against task criteria and possibly making broad revisions, Amabile, 1983, 1988; Mumford et al., 2002; Lonergan et al., 2004; Amabile & Pratt, 2016) or elaboration stage (refining and adding detail to idea; Mainemelis, 2010). If the data that the creation receives as input are seen as safe, this creates trust that the AI will be unlikely to cause adverse effects, such as reinforcing harmful biases. To illustrate, Kevin reflected on witnessing colleagues unknowingly introduce bias into their creation:

> These researchers had trained the model, and just using the data that was available, they introduced sexism and racism into a model because the data that they were provided is coming from … our world, and therefore it is racist and sexist. And thereby we are teaching it what we in a sense already know, which is those intrinsic attributes. And it's unsettling because, at least with humans, we to some degree expect that. We know what's on the table with humans. But now we're introducing something that we don't understand, and we're teaching it the same things that we have. With AI, you wouldn't necessarily expect that you'd have the same biases in there. And it's upsetting that in our quest to create a better a world with technology, you're mimicking the one that we already live in.

He described his concern about biased data and his conclusion that extra steps should be taken at the data input stage to prevent this issue:

> There's a bit of a myopic view in tech. And you know, not everybody took the liberal arts 101 classes to educate them that, you know, they are – they have these shortcomings, and that's why they don't necessarily even see the shortcomings in the data. You have to … be aware of the bias implicit in yourself in order to see it in the world around you. And, so, I think it has to start with the people who are working on it.

Ray reflected on the role that AI developers may play in developing practical solutions that prevent the AI from reinforcing unwanted biases:

> And I don't think there's any more an excuse, that, "Hey, that's in the data, so what can we do?" No, we can do things about it. And it's good to see that those kinds of systems are happening. And it's a challenge. Indeed it doesn't happen on its own, I mean the systems are biased if you just let them do what happens naturally. So we have to build technology that prevents that from happening.

Hence, some informants recognized that rather than simply using any available data, they could create a more robust tool by introducing additional steps into data processing. While it may not be possible to eliminate bias entirely, given that human bias is often implicit and unconscious (Greenwald & Krieger, 2006), these informants worked to reduce bias as much as possible and developed trust that their creations would be reliable and safe contributions to society.

Constraining one's own reactions refers to replacing enthusiasm with skepticism. This constraint takes place when a particular idea is being validated, or checked against the task criteria (Amabile, 1983, 1988; Amabile & Pratt, 2016). Informants who constrained their reactions spoke of the importance of being skeptical of positive outcomes, which they learned could be false positives and could cause harmful effects if trusted. They also described learning to be humble about what their AIs could do rather than blindly trusting that they would benefit humanity. For example, Gary described the risk of releasing a malfunctioning creation:

> Models, they're just learning from data, so sometimes they can learn from data in a way which doesn't mean that they're learning these representative, interesting concepts, but instead they're learning very particular, precise details from that particular data set. …So if you deploy that model in the real world, …then it's going to perform worse."

He said that "a lot of problems" can arise from relying solely on model output to gauge an AI's safety and usefulness once implemented in real-world settings: "By just looking at the accuracy, I misunderstood what makes a model good. And what makes a model good is that it can generalize and can be applied to similar data and give a good result." After repeated experiences with initial results misrepresenting how well the technology was actually performing, Gary learned to question positive results in order to reduce the likelihood of releasing a malfunctioning creation:

> I've kind of built a certain skepticism towards ideas and kind of trying them out. So for me it was almost like a series of times where I kind of evaluated something

> using these metrics in an incorrect way, and what ended up happening was, I evaluated it in an incorrect way, got excited, "Oh wait, let me just look into this a little bit more. I did it all wrong. I need to do it again!" So, yeah, I'd say that's where that skepticism comes from. I do something, get excited, "Oh, it didn't work, I need to be a little bit more mindful and skeptical when it comes to results." …This is a long-time build-up of: I try something, I'm so happy it works, I look into it a little bit more, it doesn't work. That happens over and over and over again until you realize, "Okay, I've got good results – why do I have good results?" Your first instinct isn't like, "Yeah!" It's, "This is weird, I'm not used to seeing this. We should look into this a little bit more." What usually happens is, okay, we did this wrong, and we did this wrong. So that means maybe these results aren't as good as we thought.

Because of the harmful outcomes that can emerge if false positives are interpreted as valid indicators, learning to be skeptical of positive outcomes is an important step in developing a robust tool that is less likely to malfunction.

Yet constraining one's own reactions involves more than the day-to-day evaluation of output. Informants also spoke of the dangerous tendency to overestimate the positive contributions of AI and to develop blind trust in AI tools. To counteract this, they also constrained their reactions by maintaining a skeptical, modest view about what their work would ultimately contribute once implemented. For example, Nelson described his perspective of maintaining humility about progress:

> Anytime engineers or someone else come up with a new technology, we tend to not see the downsides, I mean we tend to overestimate the positives. So it's not only for cynical reasons, it's because that's how we work: We want progress. We want to focus on the positive, and we need to focus on that. So, I think it is good to be humble in a sense. So, it's just kind of a reminder. You have to be humble and realize that it's very unlikely that deep neural nets will solve all or major part of all AI problems. It's going to become a more complex thing.

By approaching positive results with skepticism and maintaining a humble stance about the future of their work, these informants gained deeper knowledge of their creations' underlying functionality – and perhaps most importantly, their limitations. Informants implementing this constraint reminded themselves to question positive results, shine light on black boxes, and counteract blind trust. With this greater knowledge of their creations' capabilities and limitations, informants seemed to gain confidence that they could predict how their tools would perform, which would give them opportunities to make their creations more robust and therefore safer for the general public.

Finally, constraining the release process, a late-stage constraint that follows outcome assessment (deciding whether to move forward with implementing an idea), entails introducing extra checks into the process of implementing the creation. Informants who constrained their release process described the need for thorough testing and refinement based on the potential harm that could arise from releasing a malfunctioning creation. For instance, Christopher recounted witnessing a new AI tool fail to identify a person of color; the AI had been trained and tested only on Caucasian individuals. He recalled, "Wow, that person walked away maybe a little confused, and maybe they didn't understand why it wasn't working for them. Or maybe they got it, too, and maybe this has happened more than once for them." This experience attuned him to the harmful outcomes that may emerge from implementing a malfunctioning creation, and he developed more thorough release process. He explained, "So, yeah, and that made me think, like, 'Wow, okay so you should always have more people's eyes.' Whatever you build then goes out into the world. People are interpreting it and trying to make sense of it. And more than sharing data with it, but they're interacting with it too." Rather than proceeding to implement a tool that seemed to perform well, Christopher built more thorough checks into his release process to mitigate the risk of releasing a poorly performing AI in the future:

> …Maybe talking about it, talking about how your algorithm works, or like demoing it early so you can see the risks or perils… So I think always thinking about those risks early on prevents you from having the big hiccups. …I guess inspecting the model or, like, you could almost, like, interview this graduate, or interview this model: 'How do you do for these edge cases?' And if someone that we were selling our machine learning model to asked about those, I would really

> want to try to find some example data and see how it does. And to see, like, how does it perform for these small samples and develop a trust for that.

As with constraining input and reactions, constraining the release process involves improving the robustness of the AI creation. This allowed informants to trust that their creations would be high quality and therefore less likely to cause harm by malfunctioning.

Engagement in work: Perhaps surprisingly, informants who constrained their work ultimately seemed highly involved in, satisfied with, and enthusiastic about their creative AI work; as such, they may be described as engaged in their work (Harter, Schmidt, & Hayes, 2002). For instance, Barry described finding his work fun, challenging, and exciting:

> I think the whole process of designing the topology and stuff is a really fun experience. Like trying to come up with a neural net that hasn't existed before, like working on data and working on a problem that hasn't been solved is really exciting. …It's like, it's a new problem, and you don't know how good you can solve it, or if you can even solve it at all. That would be my takeaway message, is that there's lots of cool problems to solve and it's fun to try to solve them. I was being a little pessimistic [earlier], but there are lots of cool things you can do with it. It's actually really, really exciting stuff to be working on.

These informants also described enjoying the surprises that emerged from their work, despite their more cautious approach to the creative process. To illustrate, Steven, who imposed constraints to create a sense of both control and trust in quality, recounted his interest and amazement when working with an algorithm that produced surprising output:

> With my Ph.D. student, [name], we sat down a couple of days, we devised an algorithm, we designed an algorithm, and we're like, "Okay. Let's see how it solves a robot navigation task," without really having high hopes, right? So, we thought that if we gave it the ability of surprise, the robot, it would do interesting things, and it actually did wild, interesting things – sort of actually solved the problem much faster than if you give it a particular objective. …These are the moments that keep you going.

Imposing constraints therefore does not seem incompatible with experiencing enjoyment, interest, and even thrill in surprises.

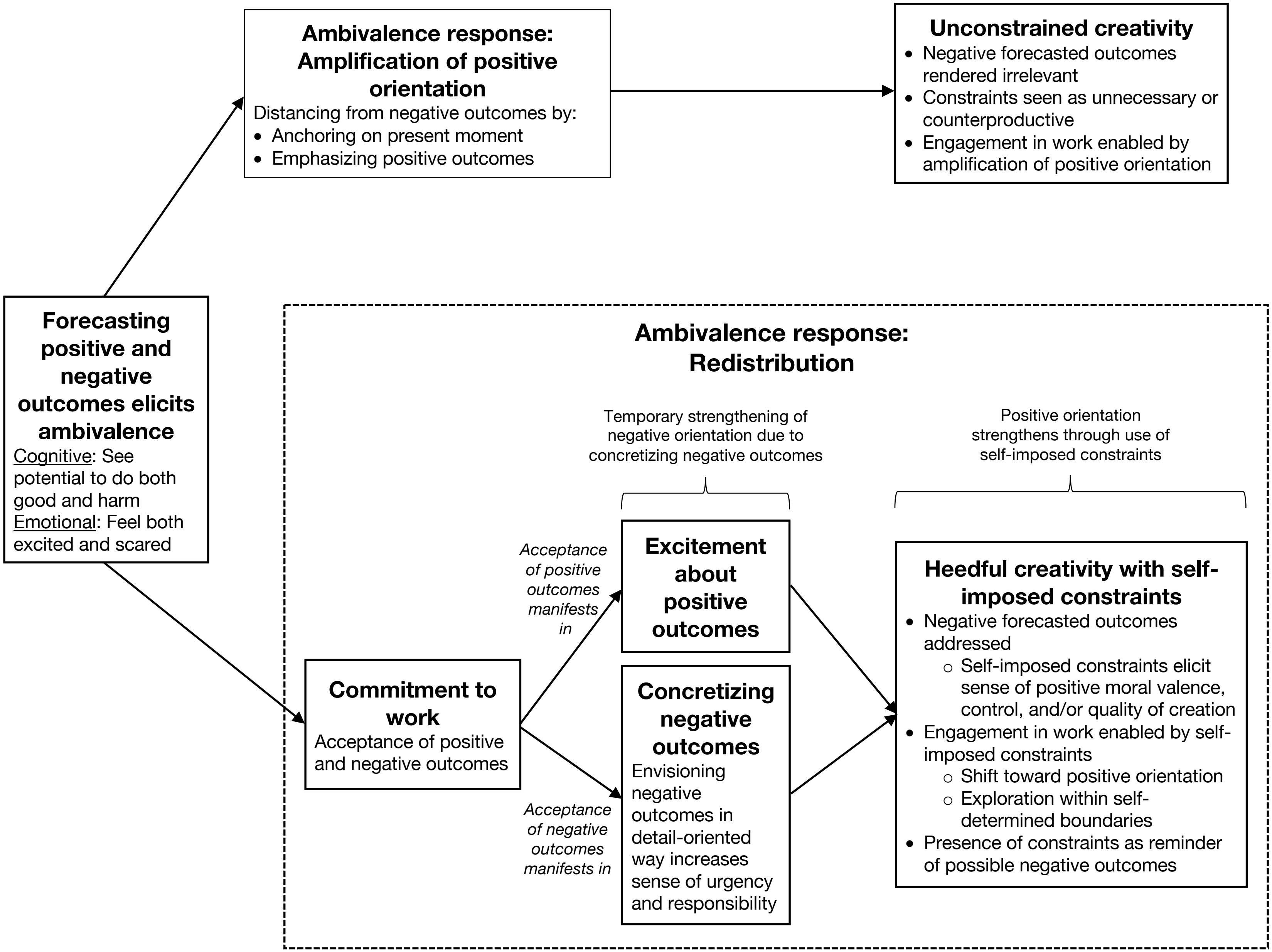
On the contrary, applying self-imposed constraints seemed to enable informants who felt responsible for negative outcomes to engage in their work more deeply than they would otherwise. Josh, who constrained the functionality of his creation, described meeting his goals to do "amazing" things with AI and to find solutions that would reduce fears:

> I think we are leaning toward that you can either be scared of this, or you can think, "I can improve the performance of this method and do something about the part that causes me to fear." So if you ask a normal person, they will say that, "Oh it's scary, I fear the implementation of the algorithm." But for us who are doing research in this context, I think we are leaning toward, "Isn't that amazing? Let's work on that, let's see how it works, and let's improve the algorithm, and let's find a solution."

Similarly, Jeremy explained, "I want to be part of this AI generation. But I also want to do it properly." He achieved this dual goal by imposing constraints that created a sense of positive moral valence and control over his creation. Ellen, who imposed project choice constraints on her work, expressed her dedication to reaping the benefits of AI by moving forward "carefully and thoughtfully." Self-imposed constraints therefore seemed to enable informants to address their concerns about the future of their work, reduce their amplified negative emotions, and ultimately feel more positive about their work.

CHAPTER 5:
DISCUSSION, CONTRIBUTIONS, AND FUTURE RESEARCH

Forecasting the future of one's creation is an inherently "noisy" process, fraught with ambiguity and uncertainty (Byrne, Shipman, & Mumford, 2010). This dissertation explores this process in the context of AI development, in which individuals grapple with the negative downstream effects that could arise from implementing their work. My analysis indicates that, in a setting in which creative ideas have the potential to harm a large number of people, individuals forecast both positive and negative outcomes and experience ambivalence. They respond to this ambivalence in different ways, which have important implications for the creative process. The induced process model encompassing these relationships follows:


Ambivalence response: Amplification of positive orientation
Distancing from negative outcomes by:
• Anchoring on present moment
• Emphasizing positive outcomes
Unconstrained creativity
• Negative forecasted outcomes rendered irrelevant
• Constraints seen as unnecessary or counterproductive
• Engagement in work enabled by amplification of positive orientation
Forecasting positive and negative outcomes elicits ambivalence
Cognitive: See potential to do both good and harm
Emotional: Feel both excited and scared
Ambivalence response: Redistribution
Temporary strengthening of negative orientation due to concretizing negative outcomes
Positive orientation strengthens through use of self-imposed constraints
Acceptance of positive outcomes manifests in
Excitement about positive outcomes
Commitment to work
Acceptance of positive and negative outcomes
Acceptance of negative outcomes manifests in
Concretizing negative outcomes
Envisioning negative outcomes in detail-oriented way increases sense of urgency and responsibility
Heedful creativity with self-imposed constraints
• Negative forecasted outcomes addressed
o Self-imposed constraints elicit sense of positive moral valence, control, and/or quality of creation
• Engagement in work enabled by self-imposed constraints
o Shift toward positive orientation
o Exploration within self-determined boundaries
• Presence of constraints as reminder of possible negative outcomes

Toward a Theory of Forecasting, Ambivalence, and Self-Imposed Constraints in the Creative Process

To delineate my emergent theory of forecasting, ambivalence, and self-imposed constraints, in this section, I summarize my findings and discuss the theoretical insights that emerge from them. To begin, I find that in the context of potentially harmful creative work, forecasting, defined as predicting the outcomes of new ideas within particular settings (Kettner, Guilford, & Christensen, 1959; Wilson, Guilford, Christensen, & Lewis, 1954; Byrne, Shipman, & Mumford, 2010; Berg, 2016) elicits ambivalence. Following current research, I define ambivalence as "simultaneously positive and negative orientations toward an object" (Ashforth, Rogers, Pratt, & Pradies, 2014: 1454) – "orientations" refers to both thoughts and feelings, and in the present research, the "object" is individuals' artificial intelligence (AI) creations, including their potential outcomes.[6] Because individuals strive for consistency (Festinger, 1957; Heider, 1958), ambivalence is an uncomfortable psychological state that individuals are naturally driven to address, which they may do in a number of ways, such as through the amplification of one orientation or through accepting both orientations (see Pratt & Pradies, 2011, and Ashforth, Rogers, Pratt, & Pradies, 2014, for overviews of different ambivalence responses).

Though all informants discussed the potential positive effects of their work, they also discussed the harmful effects associated with implementing their creations. As informants forecasted both positive and negative outcomes, they experienced cognitive and emotional ambivalence: They considered the potential to do both good and harm through their work (cognitive ambivalence) and experienced a mixture of excitement and fear (emotional

[6] While ambivalence can vary in intensity and salience and is not always a conscious experience, my data speak to ambivalence that is at least moderately intense and salient to individuals.

ambivalence). However, informants responded to this ambivalence in different ways, which are associated with different ways of relating to forecasted outcomes and with the presence or absence of self-imposed constraints.

One group of informants responded to ambivalence by anchoring on the present moment and/or the positive outcomes that they expected to emerge from their work. The former involves focusing on day-to-day work, while the latter entails emphasizing the importance of forecasted positive effects, despite potential costs that might be incurred. These informants created psychological distance (Lewin, 1951; Trope & Liberman, 2003) from forecasted negative outcomes and rendered them irrelevant to their work. These informants were enthusiastic about their day-to-day work and/or about precipitating positive outcomes through their work, and they perceived constraints as unnecessary or as impediments to their progress. As such, they proceeded without imposing constraints on their work.

This pattern aligns with prior research on "domination" responses to ambivalence, which involve amplifying thoughts and feelings of one valence over those of the opposing valence (Harrist, 2006; Bell & Esses, 2002; Katz & Glass, 1979; Ashforth et al., 2014). My work indicates that, as creative workers distance themselves from the potential negative outcomes of implementing their creations, they absolve themselves of responsibility for the negative outcomes that they could precipitate. They thereby amplify their positive thoughts and feelings toward their work and see no need to constrain their work. My work further suggests that this amplification of positive thoughts and feelings supports engagement, defined as involvement, satisfaction, and enthusiasm (Harter, Schmidt, & Hayes, 2002), in potentially harmful creative work. This path therefore validates existing research on domination responses as enabling individuals to resolve their ambivalence (Baumeister, Dale, & Sommer, 1998; Bell & Esses, 2002; Katz & Glass, 1979) – and, when the positive orientation is emphasized, to move forward with the current plan of action without cognitive dissonance (Festinger, 1957; Harmon-Jones, Harmon-Jones, & Levy, 2015). Indeed, positive emotions are associated with a sense of progress toward achieving meaningful goals (Amabile & Kramer, 2011; Higgins, Shah, & Friedman, 1997; Dweck & Leggett, 1988). The positive emotions that arise from individuals' own creations likely convey a sense of moving in the "right direction," propelling individuals to move forward with their idea development – without responding to forecasted consequences. Hence, the amplification of positive thoughts and feelings involves a dismissal of negative forecasted outcomes. This manifests in viewing constraints as unnecessary or counterproductive and moving forward without placing boundaries around the creative process or its outcomes. I therefore refer to this path as entailing "unconstrained creativity."

However, informants who concretized the potential negative outcomes of their work and employed self-imposed constraints – the majority of my sample (46 of 64) – exhibited a novel response to ambivalence, which I refer to as "redistribution." These informants accepted the positive and negative outcomes that they forecasted as likely emerging from implementing their AI creations. They thereby exhibited commitment to their work, defined as binding together positive and negative orientations through choice (Brickman et al., 1987; Pratt & Rosa, 2003; Pratt & Pradies, 2011). Following this conceptualization, commitment stabilizes human behavior (e.g., continue working in AI development) when an individual might otherwise be driven to behave differently (e.g., leave AI development and pursue other work). Commitment may be seen as having two "faces" (Brickman, 1987), in that the positive or negative orientation may, at times, be more salient to the individual, depending on the situation.

In the present case, informants' commitment manifested in both their excitement about the potential positive outcomes of their work and their concern about potential adverse effects. They faced the seemingly conflicting motivations to move forward in their work while confronting its potential consequences. Informants achieved this two-sided goal of embracing the positive aspects of their work while taking responsibility for forecasted negative outcomes by imposing constraints on their work, which enabled them to explore freely within self-determined boundaries. I refer to this process as a "redistribution" of ambivalence because this combination of commitment and self-imposed constraints seems to enable a shift from an intensification of negative thoughts and feelings toward a strengthening of positive thoughts and feelings. Below, I describe each step in this induced process.

Though individuals may be simultaneously excited and worried about the range of possible outcomes of their work, negative thoughts and emotions tend to be more impactful on behavior than positive thoughts and emotions (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001). As such, redistribution entails first a strengthening of the negative orientation. Informants concretized the negative outcomes of their work – that is, they envisioned potential adverse effects in a detail-oriented way, which seemed to temporarily amplify negative thoughts and feelings. By rendering abstract fears concrete, concretizing appears to transform fears into seemingly known consequences, reducing psychological distance (Lewin, 1951; Trope & Liberman, 2003) to negative outcomes and motivating action (Liberman, Trope, McCrea, & Sherman, 2008; McCrea, Liberman, Trope, & Sherman, 2008) to address them. Two mechanisms likely drive this effect. First, as noted above, more concrete construals of objects and events are associated with a reduction of psychological distance (Liberman, Trope, & Stephan, 2007), which drives individuals to take action. Second, negative emotions tend to arise when envisioning negative outcomes (e.g., Kahneman & Tversky, 1984), and individuals are naturally driven to reduce negative emotions (Taylor et al., 1998; Elster, 1999; Lazarus, 2001; Lazarus & Lazarus, 1994). As concretizing triggers a temporary, relative amplification of negative thoughts and feelings, this seems to exert an important influence on behavior: Individuals may shift into the analytical, critical-thinking mode that arises from negative emotions, prompting a reevaluation of the current plan of action (Clore & Huntsinger, 2007; Schwarz, 2011). In the context of potentially harmful creative work, this manifests as taking responsibility for the potential negative outcomes of implementing one's creation, as individuals are compelled to reduce perceived risks and thereby alleviate their negative emotions.

My work suggests that, because these negative thoughts and feelings are tied to creative workers' own creations – projects over which they have some degree of psychological ownership (Rouse, 2016; Pierce & Jussila, 2011; Pierce, Kostova, & Dirks, 2001, 2003) and in whose future they are invested – creative workers reevaluate the current plan of action by imposing constraints on their own work. Self-imposed constraints manifest as boundaries around the creative process or its outcomes, and they provide a sense of reducing the likelihood of forecasted negative outcomes. Individuals who apply such constraints seem to feel that they have taken responsibility for the potential negative outcomes of their work – that they have "paid their dues" – and may therefore engage in their work within these self-determined boundaries.

The redistribution response builds on – but ultimately differs from – commitment as a response to ambivalence (Brickman et al., 1987; Pratt & Rosa, 2003; Pratt & Pradies, 2011). With the redistribution response, individuals accept the positive and negative sides of their work, but unlike commitment (or other "holism" responses, Ashforth et al., 2014), positive and negative thoughts and feelings are not equally dominant. Rather, this response redistributes – or shifts the relative strength of – positive and negative thoughts and feelings. Through the application of self-imposed constraints, the temporary strengthening of negative thoughts and feelings elicited by concretizing negative outcomes shifts toward a strengthening of positive thoughts and feelings. Yet neither orientation is eliminated entirely during this process; both are important to the story.

My categorization of self-imposed constraints indicates that each type of constraint may serve one of three purposes: eliciting a sense of positive moral valence (constrain project choice), control over one's creation (constrain functionality and/or require full understanding), or trust in the quality of one's creation (constrain input, own reactions, and/or release process). Implementing constraints with these purposes seems to enable individuals to manage the negative, fear-based emotions that arise through concretizing the potential negative outcomes of their work. To explain further, if individuals believe that their creations have a positive moral valence – that creations are inherently good or safe – this alleviates fears that one's creation could be used in dangerous ways. If individuals use constraints in service of maintaining full control over their creations, this reduces the fear that their creations could harm the general public, either directly (dangerous applications, malfunctioning creation) or indirectly (job displacement), because they trust that they – or other humans – will maintain oversight and play a role in the AI's behavior. Finally, if individuals forecast risks associated with releasing a malfunctioning creation and constrain their work to increase its quality, this builds trust that the creation will be a positive contribution to the world due to its robustness as a tool, sometimes regardless of human involvement in its functionality.

Taken together, these processes show how individuals may respond to their ambivalence about their potentially harmful creations by shifting the valence of their thoughts and feelings (from more negative to more positive) through the use of self-imposed constraints. Self-imposed constraints seem to enable individuals to feel less ambivalent about their work and to engage in the creative process with enthusiasm and openness – within the boundaries of the constraint. Hence, individuals may persist in potentially harmful creative work while taking responsibility for risks. The redistribution response may therefore enable even highly concerned individuals to reap the psychological benefits of engaging in creative work, such as enjoyment, thrill, and fascination.

Yet the application of constraints also serves as a reminder of the possible negative future, should individuals decide to stop constraining their work. Informants recognized their critical role in ensuring that certain boundaries were implemented and maintained. Indeed, because such constraints are consciously self-imposed, individuals are likely aware of the threat that would reemerge without them. As such, self-imposed constraints do not resolve ambivalence entirely. Rather than eliminating negative thoughts and feelings altogether, self-imposed constraints may quell concerns and enable individuals to enjoy their work, shifting from one "face" of commitment (Brickman et al., 1987; Pratt & Rosa, 2003; Pratt & Pradies, 2011) to another. Yet the presence of self-imposed constraints serves as a reminder of the negative outcomes that could emerge without them: *As long as self-imposed constraints are used*, the positive face of commitment may face forward. As such, redistribution may be considered a *response* to ambivalence rather than a way of *resolving* ambivalence.

Because self-imposed constraints seem to enable individuals to move forward in their work while responding to the negative outcomes that they forecast, I refer to this path as involving "heedful creativity." My work builds on prior research on heedful behavior, defined as approaching situations with greater care, critical thinking, and conscientiousness (Ryle, 1949;

Weick & Roberts, 1993). Hence, the notion of "heed" describes a particular intention behind behaviors. Whereas habitual behaviors replicate past actions, heedful behaviors arise from an openness to continual learning and involve a responsiveness to the full range of outcomes that could emerge in a given situation, such as potential adverse effects. Heedfulness enables individuals to counteract the natural tendency to resist change once one has decided to pursue a particular plan of action (e.g., Brickman, Perloff, & Seligman, 1987; Brehm & Leventhal, 1962; Thibaut & Ross, 1969; Walster & Prestholdt, 1966). In settings that require ongoing alertness, heedfulness plays an important role in ensuring that operations run smoothly and that risks are appropriately addressed, which ultimately reduces organizational errors (Weick & Roberts, 1993). My work therefore translates the notion of heedfulness into the domain of potentially harmful creativity, showing how heedfulness manifests in the application of self-imposed constraints.

A note on self-imposed constraints and creative performance. Though the primary goal of this work is to show how forecasting and ambivalence influence the creative process rather than to evaluate creative outcomes, my analysis of the different types of constraints that informants imposed on their work provides tentative insights regarding how such constraints may support creativity. Building on prior research that shows how heedfulness enables individuals to attain better outcomes (Weick & Roberts, 1993), my work suggests that self-imposed constraints may support creative performance in two key ways.

Project choice, input, and functionality constraints themselves require creativity; informants developed novel ways of meeting the technical challenges before them as they worked within the boundaries set by these three constraints. As such, these constraints may be seen as redefining the problem or task criteria. To begin, project choice constraints manifested as moving away from projects for which individuals forecasted dangerous applications and toward projects that seemed inherently good or safe. Some informants made this move with the intent of focusing on a narrow range of subsequent research questions that they perceived as essential to developing safer AI. For instance, Brian explained that, though he could have "gone in many types of directions in AI," he funneled his passions for understanding human intelligence and creating AI into projects that would put "humans first." Motivated by this "continual thrust," he developed tools with the intention of augmenting and empowering humans, rather than competing with or replacing them. His work has resulted in one of the most ubiquitous "virtual assistants" currently in use. Similarly, Adam recounted constraining his project choice to focus specifically on developing AI safety techniques. He harnessed his fascination with "digging into really complicated, interesting, intellectual ideas" to tackle some of the most challenging research questions facing AI today, including ensuring that AIs act ethically. Constraining project choice may, on its surface, seem like an inherent reduction of creativity in its diminishing of the pool of ideas that one may consider. However, project choice constraints may drive creative performance by harnessing individuals' intrinsic motivation, domain-relevant skills, and creative thinking skills and applying them to challenging problems that individuals feel passionate about solving.

Constraining input, a quality constraint, manifested most often as developing creative ways of addressing the dilemma of biased AI. Informants explained that, although AI itself is neutral, it receives data from humans, and humans are biased. Informants acknowledged that eliminating bias altogether may not be possible; Greg predicted, "There's always going to be bias." Nonetheless, they viewed this as a challenging technical problem. For instance, Ray said, "[Bias] is actually a problem that we can probably solve technically. Once we recognize that this is happening, we can make sure that it doesn't happen, or improve the system. So that's something where AI researchers can be aware of that and make a difference." Informants described finding clever ways of reducing bias in their creations, such as through developing code that made it easier to locate and reduce bias. One informant, Harry, noted that although bias reduction techniques often produce bottlenecks that slow down the AI, his AI actually performed more efficiently after incorporating his novel bias-targeting function. Hence, constraining input is itself a challenging problem that requires creativity, and the solutions that individuals develop may yield AIs that are not only less biased but also perform more efficiently than expected.

The control constraint of restricting the functionality of the creation also calls for creative ideas, as informants needed to determine how and when to require human involvement or otherwise "safeguard" the AI by narrowing its range of behaviors. Like project choice and input constraints, constraining functionality is not straightforward; as Caleb said regarding the controllability of AI creations, "There's this huge debate…in AI about how we do it." Informants nonetheless came up with innovative ways of developing their creations while restricting their functionality. Kyle described finding a technique to limit the speed with which his robot moved so that it would not move in a dangerous way when humans were nearby. Bruce discussed coming up with a data-efficient technique that combined human involvement with sophisticated safety criteria to limit his swarm robots' behaviors, depending on characteristics of the environment, to reduce adverse effects. Finally, Alan recounted his evolution toward "engineering slowness" into his game design AI and building in a negotiation process between the AI and the human. He concluded, "It's made every aspect of the system better, and it's also pointed to ways that I can make healthier AI systems in the future," and he described his plan to develop additional systems that depended on humans.

Though constraining project choice, input, and functionality redefine the task criteria and drive individuals to take on particular challenges, the three remaining constraints – constraining own reactions, requiring a full understanding, and constraining the release process – may relate to creative outcomes in a different way. These constraints may benefit creativity by catalyzing the development of domain-relevant skills (Amabile, 1983, 1988; Amabile & Pratt, 2016). Domain-relevant skills involve subject matter expertise and technical skills, and they play an essential role during the task preparation and validation stages of the creative process (Amabile, 1983, 1988; Amabile & Pratt, 2016; Hirst, Van Knippenberg, & Zhou, 2009). Informants who constrained their reactions described learning to respond to positive results with doubt (Gary), trying not to "overestimate the positives" (Nelson), and having a healthy skepticism of "black boxes" in AI creations (Daniel). Constraining their reactions prompted them to double-check their work and to develop a better understanding of their creations. Informants who required a full understanding of their creations before moving forward with implementation spoke of "making the internals of the model more clear" (Gary), understanding their creations' "shortcomings" and "weaknesses" (Alan), and getting "more insight into what our current agents are actually doing" (Adam), which they deemed one of the major challenges of working with AI. Finally, informants who constrained their release process described "pilot[ing] tools on ourselves first" (George), testing new AIs on "edge cases" (i.e., unusual situations; Christopher), learning to "evaluate what's good enough" (Richard) regarding the AI's output, and working toward making creations "more correct" (Nelson). Though motivated by concern regarding possible adverse effects, these constraints propelled informants to learn more about how their creations worked and to enhance their domain skills. Because these constraints directly develop individuals' expertise, they likely support creative outcomes by boosting creative performance on subsequent tasks. Those who did not constrain their work, by contrast, sometimes described enjoying not knowing what was going on inside the "black box" of their AI creations. As such, holding all other factors equal, their creative performance would likely be lower in comparison to those who developed their domain skills due to their constraints.

Contributions and Future Research

This research makes several important theoretical contributions and opens up a number of avenues for future work. Below, I describe how this work sheds light on forecasting, ambivalence, and self-imposed constraints in creative work. I also discuss how my work aligns with and adds nuance to research on psychological distance and construal level. In addition, I discuss opportunities for future research. As an inductive, qualitative study that is focused on one domain of creative work, the strength of this project is in its deep exploration of particular theoretical dynamics in a specific context (Eisenhardt, 1989; Pettigrew, 1990). Future research is needed to test the theoretical relationships that I have induced and to explore the questions that arise from the contributions that I make.

Forecasting and creativity. One major contribution of this work is revealing how forecasting the outcomes of implementation may influence the creative process. Research on creativity has long emphasized the importance of the open exploration of ideas rather than being overly focused on outcomes (e.g., Amabile, 1979). However, in potentially harmful creative work, the possibility of precipitating adverse effects through implementing new ideas creates a tension that individuals must manage. This dissertation indicates that considering outcomes does indeed influence how creative work proceeds, but not only through increasing or decreasing intrinsic motivation (e.g., Amabile, 1979, 1985): The vast majority of individuals in my sample spoke of their deep enjoyment of their work, yet some were clearly more focused on how they could prevent potential negative outcomes than others.

My work indicates that, in the context of potentially harmful creativity, creative forecasting is not only a process of predicting whether ideas will be successful, as prior work has emphasized (e.g., Runco & Smith, 1992; Berg, 2016); it also involves considering the outcomes of implementing one's work, which may include both positive and negative effects. I find that forecasting the outcomes of implementation elicits emotional reactions that relate to whether and how individuals constrain their work in the present moment. Hence, forecasting may extend beyond considering whether task criteria will be satisfied (e.g., Lonergan, Scott, & Mumford, 2004) to include weighing the possible downstream effects of satisfying those criteria – of implementing a creative idea. This process may catalyze individuals to rethink their work altogether, such as by driving them to reject certain projects, to place boundaries on the functionality of what they are creating, or to build extra checks into the release process. Indeed, forecasting downstream effects may lead individuals to redefine the task criteria themselves, such as by changing the underlying purpose of their work. At a more general level, by showing how forecasting downstream effects seems to impact how creative workers engage with the task before them, this dissertation helps to build a bridge between creativity and innovation research: Creative workers, particularly when faced with the possibility of highly negative downstream effects, may make decisions in the present based on how they relate to the consequences of implementing their work.

Ambivalence and creativity. Though I did not enter the field searching for ambivalence, "mixed feelings" (excitement coupled with significant concerns) – associated with forecasting "mixed outcomes" – were prevalent in my data. In hindsight, this forecasting of mixed outcomes makes sense; even in low-risk situations, forecasting may not result in a singular forecasted outcome – Byrne and colleagues (2010: 120) describe forecasting as “a complex form of prediction where neither predictors nor outcomes are fixed.” My data suggest that, in the context of potentially harmful creative work, negative outcomes often seem just as likely as positive ones, and forecasting both positive and negative outcomes elicits ambivalence, defined as positive and negative thoughts and feelings (Pratt & Pradies, 2011; Ashforth, Rogers, Pratt, & Pradies, 2014), toward one’s work.

Because creativity is vulnerable to the influences of affective states (Kaufmann, 2003; Hennessey & Amabile, 2010; Amabile & Pratt, 2016), it is important to understand how ambivalence affects creative work, yet extant research on ambivalence and creativity is scant. Fong’s (2006) series of laboratory experiments indicate that ambivalence may benefit creativity by enabling individuals to develop novel associations. George and Zhou’s (2007) field study of employees in an oil field services company suggests that, in a supportive context, experiencing both positive and negative moods may drive the divergent thinking (positive moods) and more focused, critical thinking (negative moods) that are both necessary for developing novel and useful ideas. However, Fong’s work manipulated emotional ambivalence prior to – and which was unrelated to – the creative task at hand, and George and Zhou’s study evaluated affect at work using the classic Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988) and also did not relate affect to individuals’ own work output.

By showing how ambivalence may be elicited by one’s own creation and how this impacts the creative process, this dissertation also makes a major contribution to understanding the relationship between ambivalence and creativity. I show that, depending on how individuals respond to ambivalence about the positive and negative outcomes of implementing their creations, they may choose whether to impose boundaries on their work, which may affect aspects of the ideas that they develop as well as the process by which they develop and release them.

I show that, first of all, in the context of potentially harmful creative work, the "domination" response (Ashforth et al., 2014) of amplifying (Harrist, 2006; Bell & Esses, 2002; Katz & Glass, 1979) positive thoughts and feelings about one's creation is associated with moving forward in one's work without constraints. Hence, amplification of the positive orientation reduces emotional and cognitive dissonance (Festinger, 1957) and seems to support engagement in the creative process. However, it may come at the cost of paying heed to legitimate risks.

A more cautious approach to potentially harmful creativity manifests in the novel "redistribution" response, which involves excitement about realizing positive outcomes through one's work combined with a deep concern regarding negative outcomes. In a context in which individuals have some degree of psychological ownership (Pierce & Jussila, 2011; Pierce, Kostova, & Dirks, 2001, 2003) over their creations and are invested in the outcomes of implementation, these experiences appear to drive individuals to constrain their own work in various ways. Imposing constraints seems to shift an intensification of negative thoughts and feelings toward a strengthening of positive thoughts and feelings. Redistribution therefore builds on – but differs from – the commitment response to ambivalence (Brickman et al., 1987; Pratt & Rosa, 2003; Pratt & Pradies, 2011), which involves accepting both positive and negative orientations through choosing a particular object or action. To illustrate the commitment response, Brickman (1987) used the example of marriage, which involves choosing to be with a particular person whom one loves – at the cost of pursuing relationships with others. Hence, the notion of constraint is embedded in the commitment response, though Brickman did not emphasize this concept. By choosing to commit to one person, individuals accept both the positive and negative side of this arrangement. With the redistribution response, individuals commit to their work in accepting its positive and negative aspects, but concretizing negative outcomes seems to compel individuals to impose constraints on their work, which enables a shift toward positive thoughts and feelings. However, the negative orientation is not eliminated entirely through self-imposed constraints; the presence of such consciously-imposed boundaries seems to remind individuals of the harm that could arise from their work without them.

Negative emotions are associated with a narrowing of focus (Fredrickson & Branigan, 2005) that has sometimes been shown to adversely affect creative performance (Fredrickson, 1998, 2001; Hennessey & Amabile, 2010; Amabile & Pratt, 2016). Insofar as negative emotions adversely affect creativity, self-imposed constraints that address an amplification of negative thoughts and feelings may enable individuals to "recover" their creativity. Though driven by prospective thought processes, self-imposed constraints seem to enable individuals to engage more fully in the present-moment activities of exploration and invention by arming them with a sense of safe exploration.

Constraints and creativity. This work also makes an important contribution to theory on the role of constraints in creative work. I show that a variety of self-imposed constraints enable individuals to manage the concern, fear, and guilt that emerge when concretizing the negative outcomes that may arise from their creations. Much earlier research on constraints presents a largely negative perspective, focusing on how externally-imposed constraints may be detrimental to creativity (e.g., Woodman, Sawyer, & Griffin, 1993; Amabile, 1983, 1988, 1996). As an example, Woodman and colleagues (1993) defined task, time, and other resource constraints as inhibitors of creativity and placed constraints conceptually in opposition to creativity enhancers. Amabile (1988) revealed how certain types of extrinsic constraints, such as features of the social environment that are perceived as controlling individual behavior, can impede individuals' sense of freedom and autonomy (Deci, 1971, 1972; Deci & Ryan, 2002; Ryan & Deci, 2000) and reduce intrinsic motivation, ultimately adversely affecting creativity. Much research has shown that various types of externally-imposed constraints reduce creativity by curtailing engagement in the creative process (e.g., Amabile, DeJong, & Lepper, 1976; Amabile & Gitomer, 1984; Byron, Khazanchi, & Nazarian, 2010; Rosso, 2014). However, another stream of research has shown how particular types of externally-imposed constraints, such as requiring the inclusion of a particular function or feature (e.g., Finke, 1990), improve creative performance (Finke, 1990; Finke, Ward, & Smith, 1992) by counteracting the natural tendency to fall back on familiar responses (Ward, 1994; Bornstein & D'Agostino, 1992; Tversky & Hemenway, 1984; Schwarz, Bless, Strack, Klumpp, Rittenauer-Schatka, & Simons, 1991; Tversky & Kahneman, 1973). Hence, though externally-imposed constraints that reduce autonomy tend to adversely impact creative performance, externally-imposed constraints on content may serve as inputs that stimulate creative thinking.

More recently, researchers have begun to integrate different types of constraints into more unified frameworks, which help make sense of the differential effects that different types of constraints may have on creative work (e.g., Rosso, 2014; Cromwell, Amabile, & Harvey, 2018; Cromwell, 2018; Acar, Tarakci, & van Knippenberg, 2019). In Table 1, I categorized the constraints that I have induced in terms of the purpose they seemed to serve for the informants in my sample. The frameworks proposed by Cromwell and colleagues (2018) and Rosso (2014) offer particularly useful lenses through which to understand the types of constraints that I have induced and to highlight the contributions that I make. For Cromwell and colleagues (2018), "resource constraints" are similar to Rosso's (2014) "process constraints," in that these constraints limit the resources or processes used to generate ideas; these encompass constraints that have historically been seen as detrimental to creativity by reducing individuals' ability to engage in creative work. Cromwell and colleagues' (2018) "problem constraints" align somewhat with Rosso's (2014) "product constraints" and reflect how problems – and resulting creative ideas – are influenced by particular goals, such as for novelty and usefulness. Although the constraints I have induced do not map perfectly onto these two categories, Table 2 categorizes the constraints I have induced by integrating elements of these two frameworks. Additionally, I note the stages of the creative process during which they take place.

Table 2: Self-Imposed Constraints Categorized in Terms of Problem, Process, and Product Constraints

Problem Constraint	Constraining Project Choice[a]		
Process Constraints	Constraining Own Reactions[b]	Constraining Release Process[d]	Requiring Full Understanding[d]
Product Constraints	Constraining Input[b,c]	Constraining Functionality[c]	

Notes. I maintain the color scheme used in Table 1, which reflects the induced purpose of each constraint (green: positive moral valence; blue: control; light red: quality).

[a] Problem/task identification stage

[b] Idea validation stage

[c] Idea elaboration stage

[d] After outcome assessment stage

Although prior work on beneficial constraints has shown that extrinsic problem or product constraints may benefit idea generation (Finke, 1990; Finke et al., 1992; see Cromwell, 2018), this dissertation offers a different perspective, showing how multiple types of constraints – both problem/product and process – may be used proactively to manage the cognitive and emotional challenges of potentially harmful creativity. In this sense, constraints may – perhaps paradoxically – be enabling: I show that, when negative thoughts and emotions are amplified by concretizing potential negative downstream effects, there are multiple pathways toward feeling that one has taken responsibility and may move forward: Individuals may impose constraints on their own work at early, middle, and later stages of the creative process to address the types of negative outcomes that they forecast and to redistribute their ambivalence.

My findings regarding the role of self-imposed constraints in the creative process point to fruitful ground for future research. First of all, future research should test the relationships that I have induced, including whether self-imposed constraints do indeed support engagement in potentially harmful creative work by addressing ambivalence – and whether this impacts creative outcomes. As already mentioned, prior research indicates that externally-imposed constraints may adversely affect creativity by reducing autonomy and intrinsic motivation; it may be that self-imposed constraints benefit creativity not only by quelling negative emotions, but also by protecting autonomy. Given that self-imposed constraints cannot be manipulated directly, correlational field data is needed to test how self-imposed constraints influence the creative worker during the creative process. With such data, researchers should of course measure and control for individual differences that may affect creative performance, such as intrinsic motivation (e.g., Amabile, 1979, 1985).

Future research is also needed to determine whether the different types of self-imposed constraints that I have induced differentially impact creative performance, and if so, how they do so. As already noted, though my dissertation focuses on how forecasting and ambivalence influence the creative process, my work indicates that different types of constraints may support creativity in different ways: Constraints on project choice, input, and functionality require individuals to come up with novel ideas to satisfy the revised task criteria (i.e., working within the boundaries of the constraint). Constraining own reactions, requiring a full understanding, and constraining the release process likely support the development of domain-relevant skills. Though these insights are tentative, my work indicates that, in addition to helping individuals manage the challenging thoughts and emotions that arise in potentially harmful creative work, self-imposed constraints may also support creative performance.

However, these ideas should be tested, and the precise mechanisms by which different types of self-imposed constraints affect creative outcomes should be explored. The constraints that I have induced may be categorized along a number of dimensions (e.g., problem/product vs. process, early-stage vs. late-stage), each of which may be an important source of variance in terms of how they impact creative performance. For example, the early-stage constraint on project choice reduces the pool of ideas that individuals are willing to explore and thereby directly affects idea generation. Later-stage constraints, such as on the release process, may follow a more open idea generation process. Though later-stage constraints may not affect idea generation directly, they may have an indirect effect on idea generation by leaving uncomfortable emotional experiences unaddressed. That is, when potential negative outcomes are salient and negative thoughts and feelings are amplified, the stage at which this amplification is reduced may influence creative performance, given the importance of positive affect for idea generation (Hennessey & Amabile, 2010; Amabile & Pratt, 2016). Because self-imposed constraints that are applied earlier in the creative process would presumably reduce negative emotions more quickly, such early-stage constraints might improve creative performance as compared with later-stage constraints. However, if *planning* to implement a constraint curtails negative emotions, later-stage self-imposed constraints might result in the highest creative performance by enabling open idea generation *and* diminishing negative emotions. Experience-sampling methods would be useful in teasing out the nuances of the emotions and thought processes that take place, depending on the type of constraint involved.

A broader question emerges when considering that creativity sometimes calls for being comfortable with social disapproval and running against existing norms (Baucus, Norton, Baucus, & Human 2008; Kelley & Littman, 2001). If social norms emerge around constraints, such that certain constraints are no longer self-imposed but become expected, the types of constraints that I have observed as enabling engagement in the creative process might reduce creativity because they have become institutionalized and challenge individual autonomy. Qualitative methods would be helpful to explore the impact of the institutionalization of previously-self-imposed constraints. If self-imposed constraints eventually become externally imposed and begin to quell individual creativity, a provocative question arises: Do the benefits of feeling that creations are safe outweigh the risks of limiting individual creativity?

It is also important to better understand what drives certain individuals to constrain their work when forecasting potential positive and negative outcomes and experiencing ambivalence. Informants who amplified their positive thoughts and feelings may have been driven to protect their intrinsic enjoyment of their work, perhaps viewing constraints as a threat to this enjoyment. Informants who imposed constraints on their work, though, seemed more driven by prosocial and moral concerns. These informants also expressed the belief that they could effect change through their work – that imposing constraints would affect the future of their work and perhaps the future of AI more broadly. It is likely that individual differences play an important role in whether individuals respond to ambivalence by imposing constraints. Future research should explore prosocial motivation, which involves concern for others' wellbeing (Batson, 1987); moral identity (Aquino & Reed, 2002), which is associated with imagining potential consequences and devising alternative courses of action (Johnson, 1993; Narvaez & Mrkva, 2014; Werhane, 1999; Whitaker & Godwin, 2013; Keem, Shalley, Kim, & Jeong, 2018); and self-efficacy, or a sense of one's ability to achieve desired outcomes (Bandura, 1977, 1982), as influencing whether individuals are likely to impose constraints on their potentially harmful creative work. It may also be the case that there are subtle differences in motivational processes driving the different types of constraints that I have induced. For instance, constraints that create a sense of positive moral valence may be more motivated by a strong sense of prosocial motivation and moral identity, whereas constraints that create a sense of control may be more motivated by a strong sense of self-efficacy. Future research is needed to parse out these nuances.

Contextual factors may also play an important role in informing these processes. For instance, some organizations may prioritize developing cutting-edge techniques, whereas others may be oriented toward developing new technologies with caution. Although an organization-level analysis is beyond the scope of the present research, I found – perhaps surprisingly – that those imposing constraints on their work spread fairly evenly across task types (e.g., academic research and consumer product development). Though the sample size is small, this absence of a clear pattern is likely due to the newness of AI and the current lack of norms for responding to perceived risks. An important area for future research is to identify organization-level factors that influence how individuals respond to ambivalence about the potential outcomes of their work. Indeed, even subtle contextual changes, such as wording, can influence individuals' attitudes and motivational processes (Wittenbrink & Schwarz, 2007).

Psychological distance and construal level. Aspects of the processes that I have induced complement and extend theory on psychological distance and construal level. Informants who anchored on the present moment and/or emphasized positive outcomes created psychological distance (Lewin, 1951; Trope & Liberman, 2003) between their own work and potential negative outcomes, which seemed to enable them to move forward without constraints. Some informants created this psychological distance by reasoning that they should focus on the technical details of their work and that implementing constraints was not part of their role. By displacing responsibility to others, these informants were able to proceed without constraints, without apparent conflict. For other informants, psychological distance arose from reasoning that negative outcomes would be so far away that they were simply not relevant to their own day-to-day work; these individuals created psychological distance based on the forecasted temporal distance of potential threats. Concretizing, by contrast, involved envisioning negative effects in a detail-oriented way, which reduced informants' psychological distance through creating a more intricate and grounded construal of outcomes (Liberman, Trope, & Stephan, 2007).

I show how individuals may, by either distancing or concretizing, impact the psychological distance that they perceive when forecasting the detrimental outcomes of their own work, which relates to the role of self-imposed constraints in their work. I also reveal one way that these processes unfold in real-world settings. My findings suggest that in creative work, concretizing has to do with seeing the potential of what an idea could become, and through this process, the potential downstream effects of ideas become more concrete, which seems to give rise to a sense of urgency and responsibility for forecasted negative outcomes. Distancing, by contrast, can take place through focusing on one's day-to-day work or by creating a sense of temporal distance from forecasted negative outcomes, both of which seem to reduce individuals' sense of responsibility for negative outcomes.

As a brief final point, my finding that concretizing outcomes leads to the use of constraints suggests that there may be a boundary condition to prior research indicating that more concrete construals are associated with self-control failures (Fujita, Trope, Liberman, & Levin-Sagi, 2006; Fujita, 2008). Fujita and colleagues' (2006) studies indicate that inducing a high-level construal reduces the tendency to choose immediate rewards over delayed gratification – a marker of poor self-control (e.g., Ainslie, 1975; Frederick, Loewenstein, & O'Donoghue, 2002; Thaler, 1991) – as well as to forgo discomfort, when discomfort could bring about a positive distal outcome. These studies also suggest that when individuals apply low-level construals, by contrast, they fail to exert self-control. My work indicates that, when construals involve the potential adverse effects of one's own work on other people, thinking in more concrete terms motivates individuals to exert control over their own actions by imposing constraints on their work.

Conclusion

This research provides novel insights into the experiences and processes involved in potentially harmful creative work, focusing on the domain of artificial intelligence (AI). This work illuminates the process by which creative workers choose whether to constrain their work based on how they respond to the emotions elicited by forecasting the outcomes of implementing their creations. In addition to responding to the call for research to explore how individuals developing AI think about the future of their work (Amabile, 2019), this dissertation may help individuals involved in any type of potentially harmful creativity to better understand the experiences and processes underlying their work. I show that forecasting affects decisions made during the creative process, which can drastically reshape the types and qualities of ideas that individuals develop and the process by which they are developed.

Organizations should recognize that individuals working on potentially harmful creative projects, such as AI, cloning, and gene editing, are likely experiencing ambivalence – and may be inventing the standards as they go. Organizations should therefore evaluate whether employees' constraint decisions align with organizational goals. Further, organizations may be able to encourage employees to become more heedful, such as by providing opportunities for employees to reflect on potential consequences and how they might constrain their work to address them.

Additionally, given that redistribution may be considered an ambivalence *response* rather than a *resolution*, my work also suggests that some individuals' ability to manage their ambivalence may hinge on their ongoing ability to impose constraints on their work. If individuals concretize negative outcomes, but the context prohibits self-imposed constraints, this would likely leave negative thoughts and emotions amplified. Under such conditions, individuals would be challenged to reduce their negative thoughts and feelings in other ways, which could result in abandoning their work and leaving the organization. Hence, organizations engaged in potentially harmful creative work should take stock of their employees' emotional experiences and constraint preferences, not only to see whether they align with organizational goals, but also to address the possibility of turnover.

Finally, my work contributes to an emerging discussion about the risks of unconstrained creativity (e.g., Harrison & Wagner, 2016). To date, the vast majority of creativity research has looked for ways to unleash creativity by encouraging individuals to think "outside the box" (Guilford, 1967, 1968, 1982; Runco, 2010; Simonton, 1999; Gino & Wiltermuth, 2014; Gino & Ariely, 2012). Indeed, much research on creativity has uncovered how external constraints, such as excessive bureaucracy, a lack of support for creativity, time pressure, and insufficient resources, hamper individual creativity (e.g., Amabile, 1988). However, as work has shifted away from highly bureaucratic settings that ask employees to engage in rote tasks toward increasingly boundaryless settings that are dominated by knowledge work (Boland & Tenkasi, 1995; Adler, 2001; Barley & Kunda, 2006; Powell & Snellman, 2004), it may be necessary to consider how particular types of boundaries are, under certain conditions, not only helpful but necessary. I show how, in the context of potentially harmful creative work, thinking "inside the box" may have emotional benefits in enabling individuals to approach their work more heedfully – in addition to, ideally, protecting humans from those harms.

REFERENCES

Aaker, J., Drolet, A., & Griffin, D. (2008). Recalling mixed emotions. ***Journal of Consumer Research***, 35(2): 268-278.

Abbott, B. (September 2019). 'The bells start going off.' How doctors uncovered the vaping crisis. Retrieved from https://www.wsj.com/articles/the-bells-start-going-off-how-doctors-uncovered-the-vaping-crisis-11569252950.

Acar, O. A., Tarakci, M., & van Knippenberg, D. (2019). Creativity and innovation under constraints: A cross-disciplinary integrative review. ***Journal of Management***, *45*(1): 96-121.

Adler, P. S. (2001). Market, hierarchy, and trust: The knowledge economy and the future of capitalism. ***Organization Science***, 12: 215–234.

Ainslie, G. (1975). Specious reward: A behavioral theory of impulsiveness and impulse control. ***Psychological Bulletin***, 82: 463–496.

Amabile, T. M. (1979). Effects of external evaluation on artistic creativity. ***Journal of Personality and Social Psychology*,** 37: 221-233.

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. ***Journal of Personality and Social Psychology***, 43(5): 997-1013.

Amabile, T. M. (1983). The social psychology of creativity: A componential conceptualization. ***Journal of Personality and Social Psychology***, 45: 357–376.

Amabile, T. M. (1985). Motivation and creativity: Effects of motivational orientation on creative writers. ***Journal of Personality and Social Psychology***, 48: 393-399.

Amabile, T. M. (1988). A model of creativity and innovation in organizations. In B. M. Staw & L. L. Cummings (Eds.), ***Research in organizational behavior*** (vol. 10, pp. 123–167). Greenwich, CT: JAI.

Amabile, T. M. (1996). Creativity in context. Boulder, CO: Westview Press.

Amabile, T. M. (2019). Creativity, Artificial Intelligence, and a World of Surprises: Guidepost Letter for Academy of Management Discoveries. ***Academy of Management Discoveries,*** TBD.

Amabile, T. M., Barsade, S. G., Mueller, J. S., & Staw, B. M. (2005). Affect and creativity at work. ***Administrative Science Quarterly***, 50(3): 367-403.

Amabile, T. M., DeJong, W., & Lepper, M. (1976). Effects of externally imposed deadlines on subsequent intrinsic motivation. ***Journal of Personality and Social Psychology***, 34: 92-98.

Amabile, T. M., & Gitomer, J. (1984). Children's artistic creativity effects of choice in task materials. ***Personality and Social Psychology Bulletin***, 10(2): 209–215.

Amabile, T. M. & Kramer, S. J. (2011). ***The progress principle.*** Boston: Harvard Business Review Press.

Amabile, T. M., & Pratt, M. G. (2016). The dynamic componential model of creativity and innovation in organizations: Making progress, making meaning. ***Research in Organizational Behavior***, 36: 157-183.

Anderson, N., Potočnik, K., & Zhou, J. (2014). Innovation and creativity in organizations: A state-of-the-science review, prospective commentary, and guiding framework. ***Journal of Management***, 40(5): 1297-1333.

Andersson, G. (1996). The benefits of optimism: A meta-analytic review of the life orientation test**. *Personality and Individual Differences***, 21: 719–725.

Aquino, K., & Reed, A., II. (2002). The self-importance of moral identity. ***Journal of Personality and Social Psychology***, 83: 1423–1440.

Ashforth, B. E., Rogers, K. M., Pratt, M. G., & Pradies, C. (2014). Ambivalence in organizations: A multilevel approach. ***Organization Science***, 25(5): 1453-1478.

Aspinwall, L. G. (1998). Rethinking the role of positive affect in self-regulation. ***Motivation and Emotion***, 22: 1–32.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. ***Psychological Review***, *84*: 191-215.

Bandura, A. (1982). Self-efficacy mechanism in human agency. ***American Psychologist***, *37*(2): 122-147.

Barley, S. R., & Kunda, G. (2006). Gurus, hired guns, and warm bodies: Itinerant experts in a knowledge economy. Princeton, NJ: Princeton University Press.

Basadur, M. I. N., Runco, M. A., & Vega, L. A. (2000). Understanding how creative thinking skills, attitudes and behaviors work together: A causal process model. ***The Journal of Creative Behavior***, *34*(2): 77-100.

Batson, C. D. (1987). Prosocial motivation: Is it ever truly altruistic? In L. Berkowitz (Ed.), ***Advances in experimental social psychology*** (vol. 20, pp. 65–122). New York: Academic Press.

Baucus, M. S., Norton, W. I., Baucus, D. A., & Human, S. E. (2008). Fostering creativity and innovation without encouraging unethical behavior. ***Journal of Business Ethics***, 81(1): 97–115.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. (2001). Bad is stronger than good. ***Review of General Psychology***, 5: 323–370.

Baumeister, R. F., Dale, K., & Sommer, K. L. (1998). Freudian defense mechanisms and empirical findings in modern social psychology: Reaction formation, projection, displacement, undoing, isolation, sublimation, and denial. ***Journal of Personality,*** 66(6): 1081-1124.

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An interpersonal approach. ***Psychological Bulletin***, 115(2): 243-267.

Bell, D. W., & Esses, V. M. (2002). Ambivalence and response amplification: A motivational perspective. ***Personality and Social Psychology Bulletin***, 28(8): 1143-1152.

Berg, J. M. (2016). Balancing on the creative highwire: Forecasting the success of novel ideas in organizations. ***Administrative Science Quarterly***, 61(3): 433-468.

Bethe, H. (1968). J. Robert Oppenheimer, 1904-1967. ***Biographical Memoirs of Fellows of the Royal Society***, 14: 391-416.

Biernacki, P., & Waldorf, D. (1981). Snowball sampling: Problems and techniques of chain referral sampling. ***Sociological Methods & Research***, 10(2): 141-163.

Boland, R. J., Jr., & Tenkasi, R. V. 1995. Perspective making and perspective taking in communities of knowing. ***Organization Science***, 6: 350–372.

Bornstein, R. F., & D'Agostino, P. R. (1992). Stimulus recognition and the mere exposure effect. ***Journal of Personality and Social Psychology***, 63(4): 545-552.

Bowman, S. A., & Vinyard, B. T. (2004). Fast food consumption of US adults: Impact on energy and nutrient intakes and overweight status. ***Journal of the American College of Nutrition***, 23(2): 163-168.

Brehm, J. W., & Leventhal, G. S. (1962). An experiment on the effect of commitment. In *J. W. Brehm & A. R. Cohen (Eds.),* ***Explorations in cognitive dissonance***. New York: Wiley.

Brickman, P. (1987). Commitment. In C. Wortman & R. Sorrentino (Eds.), ***Commitment, conflict, and caring*** (pp. 1-18). Englewood Cliffs, NJ: Prentice-Hall.

Brickman, P., with Abbey, A., Coates, D., Dunkel-Schetter, C., Jannoff-Bulmann, R., Karuza, J., Perloff, L., Rabinowitz, V., & Seligman, C. (C. Wortman & R. Sorrentino, Eds.). (1987). ***Commitment, conflict, and caring***. Englewood Cliffs, NJ: Prentice-Hall.

Brickman, P., Perloff, L. S., & Seligman, C. (1987). Reason. In C. Wortman & R. Sorrentino (Eds.), ***Commitment, conflict, and caring*** (pp. 19-54). Englewood Cliffs, NJ: Prentice-Hall.

Brockhaus, R. H. (1980). Risk taking propensity of entrepreneurs. ***Academy of Management Journal***, 23: 509-520.

Byron, K., Khazanchi, S., & Nazarian, D. (2010). The relationship between stressors and creativity: a meta-analysis examining competing theoretical models. ***Journal of Applied Psychology*****,** *95*(1): 201-212.

Byrne, C. L., Shipman, A. S., & Mumford, M. D. (2010). The effects of forecasting on creative problem-solving: An experimental study. ***Creativity Research Journal***, 22: 119–138.

Chiles, T. H. (2003). Process theorizing: Too important to ignore in a kaleidic world. ***Academy of Management Learning & Education***, 2: 288-291.

Chin, C. (August 2018). AI is the future – but where are the women? Retrieved from https://www.wired.com/story/artificial-intelligence-researchers-gender-imbalance/.

Clore, G. L., & Huntsinger, J. R. (2007). How emotions inform judgment and regulate thought. ***Trends in Cognitive Sciences***, 11(9): 393-399.

Costa, P. T., & McCrae, R. R. (1992). ***Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual***. Odessa, FL: Psychological Assessment Resources.

Cox, D. F. (Ed.) (1967). ***Risk taking and information handling in consumer behavior***. Boston: Harvard University Press.

Creswell, J. W. (1998). ***Qualitative inquiry and research design: Choosing among 5 traditions***. Thousand Oaks, CA: Sage.

Cromwell, J. R. (2018). Further unpacking creativity with a problem-space theory of creativity and constraint. In Guclu Atinc (Ed.), ***Proceedings for the Seventy-Eighth Annual Meeting of the Academy of Management***. Online ISSN: 2151-6561.

Cromwell, J. R., Amabile, T. M., & Harvey, J-F. (2018). An integrated model of dynamic problem solving within organizational constraints. In R. Reiter-Palmon, V. L. Kennel, and J. C. Kaufman (Eds.), ***Individual creativity in the workplace***. London: Academic Press.

Cropley, A. (2006). In praise of convergent thinking. ***Creativity Research Journal***, 18: 391–404.

Cropley, D. H. (2010). The dark side of creativity: A differentiated model. In D. H. Cropley, A. J. Cropley, J. C. Kaufman, & M. A. Runco (Eds.), ***The dark side of creativity*** (pp. 360-374). Cambridge, UK: Cambridge University Press.

Csikszentmihalyi, M. (1997). ***Creativity: Flow and the psychology of discovery and invention***. New York: Harper Perennial.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. ***Journal of Personality and Social Psychology***, 18: 105-115.

Deci, E. L. (1972). The effects of contingent and noncontingent rewards and controls on intrinsic motivation. ***Organizational Behavior and Human Performance***, 8: 217-229.

Demirci, K., Akgönül, M., & Akpinar, A. (2015). Relationship of smartphone use severity with sleep quality, depression, and anxiety in university students. ***Journal of Behavioral Addictions***, 4(2): 85-92.

Dowling, G. R., & Staelin, R. (1994). A model of perceived risk and intended risk-handling activity. ***Journal of Consumer Research***, 21(1): 119-134.

Drazin, R., Glynn, M. A., & Kazanjian, R. K. (1999). Multilevel theorizing about creativity in organizations: A sensemaking perspective. ***Academy of Management Review***, 24: 286-307.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. ***Psychological Review***, 95: 256–273.

Eby, L., Hurst. C., & Butts, M. (2009). Qualitative research: The redheaded stepchild in organizational and social science research? In C. Lance & R. Vandenberg (Eds.), ***Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences*** (pp. 219-246). New York: Routledge.

Eisenhardt, K. (1989). Building theories from case research. ***Academy of Management Review,*** 14: 532-550.

Elsbach, K. D., & Kramer, R. M. (2003). Assessing creativity in Hollywood pitch meetings: Evidence for a dual-process model of creativity judgments. ***Academy of Management Journal***, 46(3): 283-301.

Elster, J. (1999). ***Strong feelings: Emotions, addiction, and human behavior***. Cambridge, Massachusetts: MIT Press.

Feldman, M., Bell, J., & Berger, M. (2003). ***Gaining access***: ***A practical and theoretical guide for qualitative researchers.*** Walnut Creek, CA: Rowman Altamira.

Festinger, L. (1957). ***A theory of cognitive dissonance*** (vol. 2). Stanford, CA: Stanford University Press.

Feynman, M. F. (Ed.) (2005). ***Perfectly reasonable deviations from the beaten track: The letters of Richard P. Feynman.*** New York: Basic Books.

Finke, R. A. (1990). Creative imagery: Discoveries and inventions in visualization. Hillsdale, NJ: Erlbaum.

Finke, R. A., Ward, T. B., & Smith, S. M. (1992). Creative cognition: Theory, research, and applications. Cambridge, MA: MIT Press.

Fong, C. T. (2006). The effects of emotional ambivalence on creativity. ***Academy of Management Journal***, 49(5): 1016-1030.

Fong, C. T., & Tiedens, L. (2002). Dueling experiences and dual ambivalences: Emotional and motivational ambivalence of women in high status positions. ***Motivation and Emotion***, 26: 105-121.

Frankwick, G. L., Walker, B. A., & Ward, J. C. (1994). Belief structures in conflict: Mapping a strategic marketing decision. ***Journal of Business Research***, *31*(2-3): 183-195.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. ***Journal of Economic Literature***, 40: 351–401.

Fredrickson, B. L. (1998). What good are positive emotions? ***Review of General Psychology***, 2: 300-319.

Fredrickson, B. L. (2001). The role of positive emotions in positive psychology: The broaden-and build theory of positive emotions. ***American Psychologist***, 56: 218-226.

Fredrickson, B. L., & Branigan, C. (2005). Positive emotions broaden the scope of attention and thought-action repertoires. ***Cognition & Emotion***, *19*(3): 313-332.

Friedman, R. S., & Förster, J. (2001). The effects of promotion and prevention cues on creativity. ***Journal of Personality and Social Psychology***, 81(6): 1001-1013.

Friend, T. (May 2018). How frightened should we be of A.I.? *Thinking about artificial intelligence can help clarify what makes us human—for better and for worse.* Retrieved from https://www.newyorker.com/magazine/2018/05/14/how-frightened-should-we-be-of-ai.

Fujita, K. (2008). Seeing the forest beyond the trees: A construal-level approach to self-control. ***Social and Personality Psychology Compass***, 2(3): 1475-1496.

Fujita, K., Trope, Y., Liberman, N., & Levin-Sagi, M. (2006). Construal levels and self-control. ***Journal of Personality and Social Psychology***, 90: 351–367.

Gardner, H. (1993). ***Creating minds***. New York: Basic Books.

George, J. M. (2007). Creativity in organizations. ***Academy of Management Annals***, 1: 439-477.

George, J. M., & Zhou, J. (2002). Understanding when bad moods foster creativity and good ones don't: the role of context and clarity of feelings. ***Journal of Applied Psychology***, 87(4): 687-697.

George, J., & Zhou, J. (2007). Dual tuning in supportive context: Joint contributions of positive mood, negative mood, and supervisory behaviors to employee creativity. ***Academy of Management Journal***, 50(3): 605–622.

Gino, F. & Ariely, D. (2012). The dark side of creativity: Original thinkers can be more dishonest. ***Journal of Personality and Social Psychology***, 102(3): 445-459.

Goodman, L. A. (2011). Comment: On respondent-driven sampling and snowball sampling in hard-to-reach populations and snowball sampling not in hard-to-reach populations. ***Sociological Methodology***, 41(1): 347-353.

Glaser, B. & Strauss, A. (1967). ***The discovery of grounded theory: Strategies for qualitative research***. Chicago, IL: Aldine.

Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. ***California Law Review***, *94*(4), 945-967.

Gruber, H. E., & Davis, S. N. (1995). Inching our way up Mount Olympus: The evolving-systems approach to creative thinking. In R. J. Sternberg (Ed.), ***The nature of creativity*** (pp. 243-270). New York: Cambridge University Press.

Guilford, J. P. (1967). ***The nature of human intelligence***. New York: McGraw-Hill.

Guilford, J. P. (1968). ***Intelligence, creativity, and their educational implications***. New York, NY: Knapp.

Guilford, J. P. (1982). Cognitive psychology's ambiguities: Some suggested remedies. ***Psychological Review***, 89: 48–59.

Hajek, P., Phillips-Waller, A., Przulj, D., Pesola, F., Smith, K. M., Bisal, N., Li, J., Parrott, S., Sasieni, P., Dawkins, L., Ross, L., Goniewicz, M., Wu, Q., & McRobbie, H. (2019). A randomized trial of e-cigarettes versus nicotine-replacement therapy. ***New England Journal of Medicine***, 380(7): 629-637.

Harmon-Jones, E., Harmon-Jones, C., & Levy, N. (2015). An action-based model of cognitive-dissonance processes. ***Current Directions in Psychological Science***, 24(3): 184-189.

Harrison, S. H., & Wagner, D. T. (2016). Spilling outside the box: The effects of individuals' creative behaviors at work on time spent with their spouses at home. ***Academy of Management Journal***, 59(3): 841-859.

Harrist, S. (2006). A phenomenological investigation of the experience of ambivalence. ***Journal of Phenomenological Psychology***, 37(1): 85–114.

Harter, J. K., Schmidt, F. L., & Hayes, T. L. (2002). Business-unit-level relationship between employee satisfaction, employee engagement, and business outcomes: A meta-analysis. ***Journal of Applied Psychology***, 87(2): 268-279.

Hecht, D. K. (2010). Imagining the bomb. In D. H. Cropley, A. J. Cropley, J. C. Kaufman, & M. A. Runco (Eds.), ***The dark side of creativity*** (pp. 72-89). Cambridge, UK: Cambridge University Press.

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. ***Social Problems,*** 44: 174–99.

Heckathorn, D. D. (2011). Comment: Snowball versus respondent-driven sampling. ***Sociological Methodology***, 41(1): 355-366.

Heider, F. (1958). ***The psychology of interpersonal relations***. New York: John Wiley & Sons.

Hennessey, B. A., & Amabile, T. M. (2010). Creativity. ***Annual Review of Psychology***, 61: 569-598.

Higgins, E. T., Shah, J. Y., & Friedman, R. (1997). Emotional responses to goal attainment: Strength of regulatory focus as moderator. ***Journal of Personality and Social Psychology***, 72: 515–525.

Hirst, G., Van Knippenberg, D., & Zhou, J. (2009). A cross-level perspective on employee creativity: Goal orientation, team learning behavior, and individual creativity. ***Academy of Management Journal***, 52(2), 280–293.

Hogarth, R. M. (1980). ***Judgement and choice***. Chichester: Wiley.

Hogarth, R. M., Portell, M., & Cuxart, A. (2007). What risks do people perceive in everyday life? A perspective gained from the experience sampling method (ESM). ***Risk Analysis***, 27: 1427-1439.

Hogarth, R. M., Portell, M., Cuxart, A., & Kolev, G. I. (2011). Emotion and reason in everyday risk perception. ***Journal of Behavioral Decision Making***, 24(2): 202-222.

Isen, A. M. (2000). Positive affect and decision making. In M. Lewis, J Haviland-Jones (Eds.), ***Handbook of emotions*** (pp. 417–35). New York: Guilford.

Isen, A. M., & Reeve, J. (2005). The influence of positive affect on intrinsic and extrinsic motivation: facilitating enjoyment of play, responsible work behavior, and self-control. ***Motivation and Emotion***, 29: 297–325.

James, K., Clark, K., & Cropanzano, R. (1999). Positive and negative creativity in groups, institutions, and organizations: A model and theoretical extension. ***Creativity Research Journal***, 12(3): 211-226.

James, K., & Taylor, A. (2010). Positive creativity and negative creativity. In D. H. Cropley, A. J. Cropley, J. C. Kaufman, & M. A. Runco (Eds.), ***The dark side of creativity*** (pp. 33-56). Cambridge, UK: Cambridge University Press.

Janssen, O., & Van Yperen, N. W. (2004). Employee’s goal orientations, the quality of leader–member exchange, and the outcomes of job performance and job satisfaction. ***Academy of Management Journal***, 47: 368–384.

Jasper, J. M. (2010). The innovation dilemma. In D. H. Cropley, A. J. Cropley, J. C. Kaufman, & M. A. Runco (Eds.), ***The dark side of creativity*** (pp. 91-113). Cambridge, UK: Cambridge University Press.

Jick, T. D. (1979). Mixing qualitative and quantitative methods: Triangulation in action. ***Administrative Science Quarterly***, 24: 602-611.

Johnson, M. (1993). ***Moral imagination: Implications of cognitive science for ethics*** (Vol. 190). Chicago: University of Chicago Press.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. ***Econometrica***, 47: 263-291.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. ***American Psychologist***, 39: 341-350.

Katz, I., & Glass, D. C. (1979). An ambivalence-amplification theory of behavior toward the stigmatized. In W. G. Austin & S. Worchel (Eds.), ***The social psychology of intergroup relations*** (pp. 55–70). Brooks-Cole: Monterey, CA.

Kaufmann, G. (2003). Expanding the mood-creativity equation. ***Creativity Research Journal***, *15*(2-3): 131-135.

Keem, S., Shalley, C. E., Kim, E., & Jeong, I. (2018). Are creative individuals bad apples? A dual pathway model of unethical behavior. ***Journal of Applied Psychology***, *103*(4): 416-431.

Kelley, T., & Littman, J. (2001). ***The art of innovation: Lessons in creativity from IDEO, America's leading design firm***. New York, NY: Currency.

Kettner, N. W., Guilford, J. P., & Christensen, P. R. (1959). A factor-analytic study across the domains of reasoning, creativity, and evaluation. ***Psychological Monographs: General and Applied***, *73*(9), 1-31.

Kharecha, P. A., & Hansen, J. M. (2013). Prevented mortality and greenhouse gas emissions from historical and projected nuclear power. ***Environmental Science Technology,*** 47(9): 4889-4895.

Kogan, N., & Wallach, M. A. (1964). ***Risk taking: A study in cognition and personality***. New York: Holt, Rinehart, & Winston.

Kramer, M. (2014). Elon Musk: Artificial intelligence is humanity's "biggest existential threat." Retrieved from http://www.livescience.com/48481-elon-musk-artificial-intelligence-threat.html.

Langley, A. (1999). Strategies for theorizing from process data. ***Academy of Management Review***, 24(4): 691-710.

Larsen, J. T., McGraw, A. P., Cacioppo, J. T. (2001). Can people feel happy and sad at the same time? ***Journal of Personality and Social Psychology***, 81: 684-696.

Lazarus, R. S. (2001). Relational meaning and discrete emotions. In K. R. Scherer, A. Schorr, A., and T. Johnstone (Eds.), ***Appraisal processes in emotion***. New York: Oxford University Press.

Lazarus, R. S., & Lazarus, B. N. (1994). ***Passion and reason: Making sense of our emotions***. New York: Oxford University Press.

Lee, T. W., Mitchell, T. R., & Sablynski, C. J. (1999). Qualitative research in organizational and vocational psychology, 1979-1999. ***Journal of Vocational Behavior,*** 55: 161-187.

Lehman, J., Clune, J., Misevic, D., Adami, C., Beaulieu, J., Bentley, P. J., Bernard, S., Belson, G., Bryson, D. M., Cheney, N., Cully, A., Donciuex, S., Dyer, F. C., Ellefsen, K. O., Feldt, R., Fischer, F., Forrest, S., Frénoy, A., Gagneé, C., Le Goff, L., Grabowski, L. M., Hodjat, B., Keller, L., Knibbe, C., Krcah, P., E. Lenski, R. E., Lipson, H., MacCurdy, R. M., Maestre, C., Miikkulainen, R., Mitri, S., Moriarty, D. E., Mouret, J. B., Nguyen, A., Ofria, C., Parizeau, M., Parsons, D., Pennock, R. T., Punch, W. F., Ray, T. S., Schoenauer, M., Shulte, E., Sims, K., Stanley, K. O., Taddei, F., Tarapore, D., Thibault, S., Weimer, W., & Watson, R. (2018). The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. ***Neural and Evolutionary Computing***, arXiv preprint, arXiv:1803.03453.

Lewin, K. (1951). ***Field theory in social science***. New York: Harper.

Licuanan, B. F., Dailey, L. R., & Mumford, M. D. (2007). Idea evaluation: Error in evaluating highly original ideas. ***The Journal of Creative Behavior***, *41*(1): 1-27.

Liberman, N., Trope, Y., McCrea, S. M., & Sherman, S. J. (2007). The effect of level of construal on the temporal distance of activity enactment. ***Journal of Experimental Social Psychology***, 43(1): 143-149.

Liberman, N., Trope, Y., & Stephan, E. (2007). Psychological distance. In A. W. Kruglanski & E. T. Higgins (Eds.), ***Social psychology: Handbook of basic principles*** (2nd ed., pp. 353–381). New York: Guilford Press.

Locke, K. (2001). ***Grounded theory in management research***. Thousand Oaks, CA: Sage.

Lonergan, D. C., Scott, G. M., & Mumford, M. D. (2004). Evaluative aspects of creative thought: Effects of appraisal and revision standards. ***Creativity Research Journal***, *16*(2-3): 231-246.

Lubart, T. I. (2001). Models of the creative process: Past, present and future. ***Creativity Research Journal***, 13: 295-308.

Madjar, N., Oldham, G. R., & Pratt, M. G. (2002). There's no place like home? The contributions of work and nonwork creativity support to employees' creative performance. ***Academy of Management Journal***, *45*(4): 757-767.

Mainemelis, C. (2010). Stealing fire: Creative deviance in the evolution of new ideas. ***Academy of Management Review***, 35(4): 558-578.

McCrea, S. M., Liberman, N., Trope, Y., & Sherman, S. J. (2008). Construal level and procrastination. ***Psychological Science***, 19(12): 1308-1314.

March, J. G. (1994). ***A primer on decision making: How decisions happen***. New York: Free Press.

McGrath, J. E. (1982). Dilemmatics: The study of research choices and dilemmas. In J. E. McGrath (Ed.), ***Judgment calls in research*** (pp. 69-102). Beverly Hills, CA: Sage.

Miles, M., & Huberman, A. M. (1994). ***Qualitative data analysis*** (2nd ed.). Thousand Oaks, CA: Sage.

Mumford, M. D. (2001). Something old, something new: Revisiting Guilford’s conception of creative problem solving. ***Creativity Research Journal***, 13: 267–276.

Mumford, M., & Gustafson, S. (1988). Creativity syndrome: Integration, application, and innovation. ***Psychological Bulletin***, 103: 27-43.

Mumford, M. D., Lonergan, D. C., & Scott, G. (2002). Evaluating creative ideas: Processes, standards, and context. ***Inquiry: Critical Thinking Across the Disciplines***, 22: 21–30.

Mumford, M. D., Schultz, R. A., & Van Doorn, J. R. (2001). Performance in planning: Processes, requirements, and errors. ***Review of General Psychology***, 5: 213–240.

Murgia, M., & Shrikanth, S. (April 2019). How Big Tech is struggling with the ethics of AI: Companies criticised for overruling and even dissolving ethics boards. Retrieved from https://www.ft.com/content/a3328ce4-60ef-11e9-b285-3acd5d43599e.

Narvaez, D., & Mrkva, K. (2014). The development of moral imagination. In S. Moron, D. Cropley, & J. C. Kaufman (Eds.), ***The ethics of creativity*** (pp. 25–45). Basingstoke, United Kingdom: Palgrave Macmillan.

Nestler, S., Blank, H., & von Collani, G. (2008). Hindsight bias and causal attribution: A causal model theory of creeping determinism. ***Social Psychology***, 39: 182–188.

Nystrom, H. (1979). ***Creativity and innovation***. London: Wiley.

Oldham, G. R., & Cummings, A. (1996). Employee creativity: Personal and contextual factors at work. ***Academy of Management Journal***, 39(3), 607-634.

Önkal, D., Yates, J. F., Simga-Mugan, C., & Öztin, Ş. (2003). Professional vs. amateur judgment accuracy: The case of foreign exchange rates. ***Organizational Behavior and Human Decision Processes***, *91*(2): 169-185.

Pant, P. N., & Starbuck, W. H. (1990). Innocents in the forest: Forecasting and research methods. ***Journal of Management***, *16*(2): 433-460.

Patton, M. Q. (2001). ***Qualitative research and evaluation methods***. Thousand Oaks, CA: Sage.

Pereira, M. A., Kartashov, A. I., Ebbeling, C. B., Van Horn, L., Slattery, M. L., Jacobs Jr, D. R., & Ludwig, D. S. (2005). Fast-food habits, weight gain, and insulin resistance (the CARDIA study): 15-year prospective analysis. ***The Lancet***, 365(9453): 36-42.

Perry-Smith, J. E. (2006). Social yet creative: The role of social relationships in facilitating individual creativity. ***Academy of Management Journal***, 49: 85–101.

Pettigrew, A. M. (1990). Longitudinal field research on change: Theory and practice.

Organization Science, 1(3): 267-292.

Petriglieri, G., Ashford, S. J., & Wrzesniewski, A. (2018). Agony and ecstasy in the gig economy: Cultivating holding environments for precarious and personalized work identities. ***Administrative Science Quarterly***, 1: 1-47.

Pierce, J. L., & Jussila, I. (2011). Psychological ownership and the organizational context: Theory, research evidence, and application. Northampton, MA: Edward Elgar Publishing.

Pierce, J. L., Kostova, T., & Dirks, K. T. (2001). Toward a theory of psychological ownership in organizations. ***Academy of Management Review***, 26: 298-310.

Pierce, J. L., Kostova, T., & Dirks, K. T. (2003). The state of psychological ownership: Integrating and extending a century of research. ***Review of General Psychology***, 7: 84-107.

Powell, W. W., & Snellman, K. 2004. The knowledge economy. ***Annual Review of Sociology***, 30: 199–220.

Pratt, M. G., & Doucet, L. (2000). Ambivalent feelings in organizational relationships. In S. Fineman (Ed.), ***Emotion in organizations*** (pp. 204-226). London: Sage.

Pratt, M. G., & Pradies, C. (2011). Just a good place to visit? Exploring positive responses to psychological ambivalence. In K. S. Cameron & G. M. Spreitzer (Eds.), ***The Oxford handbook of positive organizational scholarship*** (pp. 924–937). Oxford: Oxford University Press.

Pratt, M. G., Rockmann, K. W., & Kaufmann, J. B. (2006). Constructing professional identity: The role of work and identity learning cycles in the customization of identity among medical residents. ***Academy of Management Journal,*** 49(2): 235-262.

Pratt, M. G., & Rosa, J. A. (2003). Transforming work-family conflict into commitment in network marketing organizations. ***Academy of Management Journal***, 46(4): 395–418.

Roese, N. J., & Vohs, K. D. (2012). Hindsight bias. ***Perspectives on Psychological Science***, 7(5): 411-426.

Rouse, E. D. (2016). In the space between: Creative workers' psychological ownership in idea handoffs. In John Humphreys (Ed.), ***Proceedings of the Seventy-sixth Annual Meeting of the Academy of Management***. Online ISSN: 2151-6561.

Rosso, B. D. (2014). Creativity and constraints: Exploring the role of constraints in the creative processes of research and development teams. ***Organization Studies***, 35(4): 551–585.

Rothman, N. B., & Wiesenfeld, B. M. (2007). The social consequences of expressing emotional ambivalence in groups and teams. In E. A. Mannix, M. A. Neale, & C. P. Anderson (Eds.), ***Research on managing groups and teams*** (vol. 10, pp. 275-308).

Runco, M. A. (2010). Creativity has no dark side. In D. H. Cropley, A. J. Cropley, J. C. Kaufman, & M. A. Runco (Eds.), ***The dark side of creativity*** (pp. 15-32). New York, NY: Cambridge University Press.

Runco, M. A., & Smith, W. R. (1992). Interpersonal and intrapersonal evaluations of creative ideas. ***Personality and Individual Differences,*** *13*(3): 295-302.

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

Ryle, G. (1949). ***The concept of mind***. Chicago: University of Chicago Press.

Scheier. M. F. & Carver, C. S. (1985). Optimism, coping, and health: Assessment and implications of generalized outcome expectancies. ***Health Psychology***, 4: 219-247.

Scheier, M. F., & Carver, C. S. (1992). Effects of optimism on psychological and physical well-being: Theoretical overview and empirical update. ***Cognitive Therapy and Research***, 16: 201–228.

Schwarz, N. (2002). Situated cognition and the wisdom of feelings: Cognitive tuning. In L. Feldman Barrett & P. Salovey (Eds.), ***The wisdom in feelings*** (pp. 144-166). New York: Guilford.

Schwarz, N. (2011). Feelings-as-information theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), ***Handbook of theories of social psychology*** (pp. 289-308). Thousand Oaks, CA: Sage.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. ***Journal of Personality and Social Psychology***, 61(2): 195-202.

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. ***Journal of Personality and Social Psychology***, 45: 513-523.

Schwarz, N., & Clore, G. L. (2003). Mood as information. ***Psychological Inquiry***, 14: 296-303.

Shalley, C. E., Zhou, J., & Oldham, G. R. (2004). Effects of personal and contextual characteristics on creativity. ***Journal of Management***, 30: 933-958.

Silvia, P. J. (2008). Discernment and creativity: How well can people identify their most creative ideas? ***Psychology of Aesthetics, Creativity, and the Arts***, 2: 139–146.

Simmons, A. L., & Ren, R. (2009). The influence of goal orientation and risk on creativity. ***Creativity Research Journal***, 21: 400-408.

Simonton, D. K. (1999). Creativity as blind variation and selective retention: Is the creative process Darwinian? ***Psychological Inquiry***, 10: 309–328.

Sincoff, J. B. (1990). The psychological characteristics of ambivalent people. ***Clinical Psychology Review***, 10(1): 43–68.

Sitkin, S. B., & Weingart, L. R. (1995). Determinants of risky decision-making behavior: A test of the mediating role of risk perceptions and propensity. ***Academy of Management Journal***, 38: 1573-1592.

Sjöberg, L. (1998). Why do some people demand risk reduction? In S. Lydersen, G. K. Hansen, & H. A. Sandtorv (Eds.), ***ESREL-98: Safety and reliability*** (pp. 751-758). Trondheim, Norway: A. A. Balkema.

Sjöberg, L. (1999). Consequences of perceived risk: Demand for mitigation. ***Journal of Risk Research***, 2: 129-149.

Sonenshein, S. (2007). The role of construction, intuition, and justification in responding to ethical issues at work: The sensemaking-intuition model. ***Academy of Management Review***, 32: 1022-1040.

Spradley, J. (1979). ***The ethnographic interview***. New York: Holt, Rinehart & Winston.

Stein, M. I. (1974). ***Stimulating creativity*** (vol. 1). New York: Academic Press.

Strauss, A., & Corbin, J. (1998). ***Basics of qualitative research: Techniques and procedures for developing grounded theory*** (2nd ed.). Thousand Oaks, CA: Sage.

Taylor, S. E., Pham, L. B., Rivkin, I. D., & Armor, D. A. (1998). Harnessing the imagination: Mental simulation, self-regulation, and coping. ***American Psychologist***, 53(4): 429-439.

Taylor, S. E., & Schneider, S. K. (1989). Coping and the simulation of events. ***Social Cognition***, *7*(2): 174-194.

Tegmark, M. (2019). Benefits and risks of artificial intelligence. Retrieved from https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/?cn-reloaded=1.

Thaler, R. H. (1991). ***Quasi rational economics***. New York: Russel Sage Foundation.

Thibaut, J., & Ross, M. (1969). Commitment and experience as determinants of assimilation and contrast. ***Journal of Personality and Social Psychology***, 13(4), 322-329.

Thomas, J. B., Clark, S. M., & Gioia, D. A. (1993). Strategic sensemaking and organizational performance: Linkages among scanning, interpretation, action, and outcomes. ***Academy of Management Journal***, *36*(2): 239-270.

Tolentino, J. (May 2018). The promise of vaping and the rise of Juul. Retrieved from https://www.newyorker.com/magazine/2018/05/14/the-promise-of-vaping-and-the-rise-of-juul.

Trope, Y., & Liberman, N. (2003). Temporal construal. ***Psychological Review***, 110(3): 403-421.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. ***Cognitive Psychology***, 5(2): 207–232.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. ***Journal of Experimental Psychology: General***, 113(2): 169–193.

Urban, T. (January 2015). The AI revolution: Our immortality or extinction. Retrieved from https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-2.html.

Vyas, K. (December 2018). 7 ways AI will help humanity, not harm it. Retrieved from https://interestingengineering.com/7-ways-ai-will-help-humanity-not-harm-it.

Wallas, G. (1926). ***The art of thought***. New York: Harcourt, Brace.

Walster, E., & Prestholdt, P. (1966). The effect of misjudging another: Over-compensation or dissonance reduction? ***Journal of Experimental Social Psychology***, *2*(1): 85-97.

Ward, T. B. (1994). Structured imagination: The role of category structure in exemplar generation. ***Cognitive Psychology***, 27(1): 1–40.

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.

Weick, K. E. (1979). ***The social psychology of organizing*** (2nd ed.). New York: McGraw-Hill.

Weick, K. E. (1995). ***Sensemaking in organizations***. Thousand Oaks, CA: Sage.

Weick, K. E., & Roberts, K. H. (1993). Collective mind in organizations: Heedful interrelating on flight decks. ***Administrative Science Quarterly***, 38(3): 357-381.

Werhane, P. H. (1999). ***Moral imagination***. New York: Wiley, Ltd.

West, M. A., & Farr, J. L. (1990). Innovation at work. In M. A. West, & J. L. Farr (Eds.), ***Innovation and creativity at work: Psychological and organizational strategies***. Chichester, UK: Wiley.

Wheelis, M., Rózsa, L., & Dando, M. (Eds.) (2006). ***Deadly cultures: Biological weapons since 1945***. Cambridge, MA: Harvard University Press.

Whitaker, B. G., & Godwin, L. N. (2013). The antecedents of moral imagination in the workplace: A social cognitive theory perspective. ***Journal of Business Ethics***, *114*: 61–73.

Williams, P., & Aaker, J. (2002). Can mixed emotions peacefully coexist? ***Journal of Consumer Research***, 28: 636-649.

Wilson, R. C., Guilford, J. P., Christensen, P. R., & Lewis, D. J. (1954). A factor-analytic study of creative-thinking abilities. ***Psychometrika***, *19*(4): 297-311.

Wittenbrink, B., & Schwarz, N. (2007). Implicit measures of attitudes. New York: Guilford Press.

Wolcott, H. F. (1994). ***Transforming qualitative data: Description, analysis, and interpretation***. Thousand Oaks, CA: Sage.

Woodman, R. W., Sawyer, J. E., & Griffin, R. W. (1993). Toward a theory of organizational creativity. ***Academy of Management Review***, 18(2): 293-321.

Zaitseva, M. N. (2010). Subjugating the creative mind: The Soviet Biological Weapons Program and the role of the state. In D. H. Cropley, A. J. Cropley, J. C. Kaufman, & M. A. Runco (Eds.), ***The dark side of creativity*** (pp. 57-71). Cambridge, UK: Cambridge University Press.

Zaltman, G., & Wallendorf, M. (1983). ***Consumer behavior: Basic findings and management implications***. Hoboken, NJ: Wiley.

Zhou, J., & George, J. M. (2001). When job dissatisfaction leads to creativity: Encouraging the expression of voice. ***Academy of Management Journal***, 44: 682-696.

Zhou, J., & George, J. M. (2003). Awakening employee creativity: The role of leader emotional intelligence. ***Leadership Quarterly***, 14: 545–568.

Zipkin, N. (March 2018). 7 weird and wild things Elon Musk said at SXSW. Retrieved from https://www.entrepreneur.com/article/310313.

Appendix I: Semi-Structured Interview Protocol for Pilot Interviews

A. Background Information [part of interview only if not yet obtained]

1. What is your educational background?
2. What is your professional background?
3. How long have you been at [organization]?
4. What is your role at [organization]?

B. Daily Work Experiences and Problem Solving

1. What are your current projects?
2. How do you decide what machine learning techniques to use?
3. What sorts of challenges typically come up in your work?
    a. How do you problem-solve or troubleshoot?
4. [Based on experiences described above] How do you come up with ideas when problem-solving or troubleshooting?
    a. How automatic is this process for you?
    b. Does emotion play a role in this process?
    c. Do you ever reflect on human processes to give you ideas for what may improve the model?
    d. Could you describe a time when you felt like you were being creative or innovative in solving a problem?

C. AI-Human Comparisons

1. What do you see as the similarities and differences between artificial intelligence and human intelligence?
2. How do you see the boundary between the AI tools you're building and the individuals who use them?
    a. In other words, what does the AI accomplish, and what is left to the person, if anything?
    b. What is the logic behind this?
3. Do you reflect on your own mental processes while working? If so, how?
4. Does working with an AI ever change the way you think about how the human mind works? If so, how?

D. Emotional Experiences and Complexity

1. Could you tell me about a time when you had an emotional reaction while working?
    a. What did you do with that (those) feeling(s)? How did you move forward?
    b. How often does this happen?
2. Have you experienced other emotional reactions in your work? Could you describe them?
    a. What did you do with that (those) feeling(s)? How did you move forward?
    b. How often does this happen?
3. To what extent do you feel that you fully understand what you are creating?
    a. Do you think about this as you do your work? If so, what emotion(s) tend to come up?
        i. What do you do with that feeling? How do you move forward?
        ii. How often does this happen?

E. Future of Work and AI in General and Emotions

1. What do you see as the overarching goal of your work?
   a. For your organization? For consumers? For society?
2. What sort of impact do you think AI is having on the world right now?
3. What sort of impact do you think AI will have in the future?
4. What emotions come up as you think about these topics?
5. Do these emotions affect how you approach your work in any way? If so, how?
   a. For instance, does your sense of the future of AI affect how you approach the problems you're trying to solve? Please explain.

F. Closing

1. Are there any other particularly salient memories that come to mind that you'd like to share?
2. Is there anything else related to these topics that you'd like to mention?

Appendix II: Final Semi-Structured Interview Protocol

A. Background/Basic Information [part of interview only if not yet obtained]

1. What is your educational background?
2. What is your professional background?
3. How long have you been at [organization]?
4. What is your role at [organization]?

B. Daily Work Experiences, Motivations, and Problem Solving

1. What initially led you to pursue work in AI/ML?
2. What are your current projects?
3. How do you decide which projects to pursue?
4. [If not covered in 1 or 3] What motivates you or drives you in your work?
    a. [Prompt if needed] For instance, people might be motivated by understanding how things work, solving hard problems, exploring ideas, inventing new things, or building tools people can use, among other possibilities. What would you say are the main motivators for you?
5. What are your typical day-to-day work experiences when developing a new algorithm?
6. Think of a recent time when you had to work through a challenge with an AI that you are (or were) developing. What types of thoughts and emotions came up for you?
    a. What did you do with that (those) feeling(s)? How did you move forward?
    b. [Prompt if needed] Did you learn anything from this experience? If so, what did you learn? If not, why not?
    c. How often do you have experiences like these? If rarely, please describe the more typical challenges that come up, if any.
7. Some interviewees have referenced the “black box” concept as applying to machine learning work. Does this ring true for you? If so, how? If not, why not?
    a. What are your views on whether machine learning models can be fully understood by humans?
        i. What leads you to think this way? Please describe any experiences that come to mind.

C. Forecasting and Constraints

1. As you’re doing your work, what sorts of possible applications or ultimate uses are you thinking about? What do you see your work building toward in the future?
    a. How do you see the relationship between humans and the technologies that you are creating?
2. [If not covered in 1] Do you ever think of more downstream outcomes that might unfold? If so, what types of outcomes do you think about? If not, why not?
    a. [Prompt if needed] For instance, imagine that your creation is very successful and widely adopted. What effects do you think that will have?
    b. Please describe any experiences that made these factors salient to you, if any.
    c. [If negative outcomes discussed] What can be done about these issues, if anything?
        i. Does your work play a role in this process? If so, how? If not, why not?
    d. Have you always held this (these) view(s)? If so, why? If not, what changed?

3. How do these considerations affect you and your work, if at all? If they don't, why not?
    a. [Prompt if needed] For instance, are you changing aspects of your work process or product based on these factors? If so, in what ways? If not, why not?
    b. [If negative outcomes discussed] Do you impose boundaries or constraints on any aspect of your work process based on these factors? If so, in what ways? If not, why not?
4. [If constraints described for 3] Did a particular experience lead you to approach your work in this way? If so, please describe it.
    a. How did you feel about your work when you *weren't* placing these sorts of boundaries on your work?
    b. How do you feel as a *result* of placing these sorts of boundaries on your work?
5. Some interviewees describe having a sense of responsibility for the possible outcomes or downstream effects of their work. Does this ring true for you? If so, how? If not, why not?
    a. [If yes] Did a particular experience make these factors salient? If so, please describe it.

D. Broader Future of AI

1. How do you see the relationship between humans and AI? For example, do you think of AI as a friend, adversary, partner, tool that humans use, or something else?
2. What do you see as the future of AI in general?
    a. How do you think society will be affected, if at all?
3. Do you think that we should have any particular concerns about how AI might develop? If so, please describe them. If not, why not?
4. If you envision a future in which AI reaches human-level intelligence or reasoning capabilities across diverse tasks, what emotions does that bring up for you?
5. How easy or challenging do you find it to predict how AI might develop over the next 10, 20, or 50 years?
    a. What leads you to think this way? Please describe any experiences that come to mind.
    b. [If relevant] When you think about the ambiguity or uncertainty involved in how AI might develop, what reactions or emotions come up?
6. Do your views on how you think AI will develop affect you and your work in any way?
    a. [Prompt if needed] For instance, do you make decisions as you're developing this technology based on these factors?
    b. Have you changed anything about your work based on these views? If so, how? If not, why not?
7. [If not covered in 5] What do you think should be done about these issues, if anything?
    a. How do you see this relating to your own work, if at all?
8. Have you always felt this way about the future of AI? If so, why? If not, what changed?
9. What current or future applications of AI are most exciting to you?

E. Closing

1. Is there anything else that you think I should know – anything related to these topics that you'd like to mention?

Appendix III: Data Table with Representative Quotations

Note: For “A Common Starting Point,” “U” indicates that an informant exhibited unconstrained creativity. “H” indicates that an informant exhibited heedful creativity.

	A Common Starting Point: “Noisy” Forecasting, Positive and Negative Outcomes, and Ambivalence
“Noisy” Forecasting	“These kinds of technologies can suddenly jump ahead much quicker than you may expect. So I still think it’s important to keep an eye out and see these AIs aren’t suddenly becoming much more intelligent than we anticipated in a short period of time.” [James - U] “Anyone you talk to, I’d say if they’re making claims about we’re approaching general human capability, in some sense, we’re just not anywhere close to that, I mean nowhere close. But at the same time, the current momentum and trajectory and velocity in the industry is just astonishing. And so in 50 years that could be completely different. In 50 years – you know, it’s actually a really short, amazingly small timescale. So we’re not there yet. Will we get there soon? Quite possibly.” [Jack - U] “Right now, the field is moving so quickly that even, you know, it’s hard to predict even in the next two years what technologies would emerge. …It’s hard to say. Ten years even seems so long. It’s hard to imagine what would be the state of the world in that sense.” [Noah - H] “Being able to estimate where research will get in 20 years is just, I mean, historically, it’s something that people try to do but never works.” [Charles - U] “Here, in my work, it’s very difficult to make this kind of [prediction] looking forward in technology because, as you may know, something may suddenly happen, and everything changes. Five years ago, deep learning was nothing, and now deep learning is everywhere. …Rodney Brooks said, ‘We’ll have limited self-driving cars in 20 years.’ And Elon Musk tells you, ‘Next year we’re going to Tesla doing self-driving cars all around.’ [Laughing] So, it’s really, really, really, really, difficult.” [Jeremy - H] “I think a lot of the time when things happen, it’s only when we look back, we say, ‘Oh, it should have happened.’ But no one actually does a good job to predict the future. …We can foresee for maybe a couple of years, and what comes after that may be totally different from what we can foresee. …That uncertainty is in our

	work. But the problems we are trying to solve is something we know; it's just how we get there we don't know." [Eve - U]	
Forecasted Outcomes	**Positive:** **Driving Scientific Progress** "I am part of that generation of AI researchers who gained perhaps initial inspiration as a child from the famous HAL in the film 2001. …And then of course as I started graduate study, PhD study, I encountered a lot of different kinds of ways into that part of computer science. I was particularly interested in evolution as a problem solver, really, but as a means to develop these intelligent systems. I have always been very interested in the idea of using inspiration from nature to try to find ways of engineering solutions to the problem of intelligence." [Ellen - H] "The real question there is really for me about how complexity can be created out of almost nothing. That's what really inspired me. So it's like, my brain is a product of evolution, which is an unguided search process. …So that's an open-ended process, it just keeps going, it keeps inventing things. It doesn't converge	**Negative:** **Causing Direct Harm Through Dangerous Applications** "In 2016, we were one of if not *the* state of the art. But just two years later - nowadays technology advances so fast, that the fake images generated by machines, by AIs, look just like real. And clearly, from my own line of work, this is very scary, because if it is not used by good people, you can do fake news, you can generate porn, and many negative things. So that is personally in my line of work, it's scary." [Bob - U] "It's an efficient and effective technology for detecting things, capturing things and then, you know, like, creating things. …So, yeah, that … you have to be worried about. Because you have this kind of effective and efficient AI technology, we can make a better as in, like, you know, worse, kind of weapons or the things that can hurt a lot of people." [Nathan - U] "The idea of weaponizing any kind of technology is of great concern. And the more that you can depersonalize doing bad stuff to other people, the easier it becomes." [Ellen - H] "These debates and tradeoffs, people think about this in any line of work, but the fact is that in ML, the impact of a single algorithm can be so far reaching because it can be applied to influence the software that so many people use, that it does give it extra weight. Someone who's debating whether to build a building for the government or not, it's just one building, but if you debate whether to build an algorithm for the government or not, that algorithm could potentially be applied to hundreds of millions of people." [Colton - H]

	to this endpoint and then it's done. I think it may be that we need an open-ended process to actually produce something like a brain. Because we are a product of an open-ended process. So trying to understand how to produce open-ended processes is important for the advancement of the field of AI and it's something that we're working on." [Ryan - H] "Honestly, in the end, sure, I want to be good and do good stuff for the world, but honestly, I mean when we talk about motivations, I care more about inventing new interesting stuff. So what really pushes me is inventing new things. That's what I see as a value in itself." [Mark - U] "We also want to focus on the main scientific question. Our main goal is not to have it next year deployed in the environment. It's to be able to find the right way of giving these kinds of capabilities to robots. ...But I will say, as a scientist, I care more about general capabilities than the specific application." [Daniel - H] "It's kind of fascinating, it's one of the things that drew me to AI, was	"I'm not 100% sure to what degree everyone having this technology will prevent it from being used in malicious ways. For example, one of the technologies that worries me at least a little bit is the ability of these neural networks to create fake images, fake videos. They're starting to become increasingly realistic, so it becomes hard to tell them apart from reality." [James - U] "They say, 'Oh yeah, we have an airplane that crashed, a similar airplane that crashed, and we need to figure out the people who get out of the plane who survived and find them and be able to rescue them with a search and rescue team.' So everything is kind of really civil security and everything... But just get the context and put that in, 'We have an enemy plane that crashed, and we want to kill them, so can we just find them in the world and fire a gun?'" [Henry - H] **Causing Direct Harm Through Malfunctioning Creations** "There is I think a much more serious and much more concrete issue of bias in machine learning systems which we definitely need to overcome. …I think machine learning needs to have much more input from social science and areas that have actually thought about these issues a lot more deeply than machine learning has. I think there is work to be done in aligning the incentives of machines with humans more closely. I don't exactly know how that would work, but I think there are interesting research directions there too. So there are some concrete needs now." [Sean - U] "Sometimes the consequence is just the company lost some money, but sometimes it could be really serious. So in this case, the people will try to develop some very correct system. They prefer to lose some detection, but once they detect something or try to predict something, it will be really very accurate. Just like self-driving. In this kind of application, it's not

	all these sorts of analogies between creating an algorithm that searches, and then ourselves, human inventors that are searching for algorithms. It’s kind of complicated and interesting.” [Adam - H] **Improving People’s Everyday Lives** “Everybody talks about, like, ‘Well is it the 4th or the 5th or the Nth industrial revolution?’ I’m like, ‘Yeah, sure. Now we can replace some other class of labor with machines. Nice! People can do something else.’ If we can replace it with a machine, it probably wasn’t worth doing with a human in the first place. Of course, this is kind of self-justifying, but it’s also what I believe. …I don’t see any big reason to think it’s going to be that different.” [Mark - U] “If I had to decide to design society, it seems like you could use this stuff and it would make for a more free and – an environment where an individual can thrive more easily.” [Frank - H]	just like a fraud detection system. There are very important consequences.” [Marie - H] “[T]he systems are not always right. So clearly there is a chance you would hit the wrong target [with autonomous weapons], and that would be very bad for humankind. That’s the first thing I worry about.” [Bob - U] “If you’re a black man, you’ve just been rejected a loan, you think this is very weird, ‘My credit score is good, okay, why?’ And they say, ‘Well, you know, the ML algorithm figured that out.’ Well what’s probably happened is there’s bias in the data, because this is sourced from the past where they were a lot more racist. And what’s happened is that it’s learned the pattern of if you’re black, we’re not going to give you a loan. And they don’t even know that’s in their model. But if you actually go in and explain it, you’re able to be a lot more fair, you can remove that whole process from the model. It’s really important for me that we’re able to understand what our models are doing.” [Gary - H] “What’s the goal? What is the end game? Do you want to make it so you don’t have to do anything? Because if that’s the end game, you’re going to fail. In the end, you’re going to end up with a system that you don’t understand, that makes predictions or recommendations based on data that you haven’t read and don’t have the same level of competence for as the network does, and now it’s going to be making decisions based on that data, and you have no way to check it.” [Barry - H] **Causing Indirect Harm Through Job Displacement** “I think the big threat is probably just the lack of employment moving forward, more of the economic effects of what happens when labor isn’t valuable anymore, because AI can basically do anything. …It’s big threat.” [Jim - U]

	"I think people always talk about this job displacement. However, if you look back in history there are three industrial revolutions. And each time – for example there is the machines replace workers. And each time you would think that there are people being replaced by something, so there is job loss. I think this is only true for a shorter period of time. There is always jobs in terms of jobs after that when you look at the history of what happened. People actually will be freed from some of the highly laborious and be able to do some more interesting work, and the work is actually more diverse." [Eve - U] "If we can teach a machine something, it would make it intelligent, so that our life would be easy, so that it can work for us." [George - H] "I want it to be useful to people. I want it to augment our own capabilities, and I want to find a way to make it improve our lives and make our lives more fulfilling." [Harry - H]	"I mean you could definitely use it as a tool, but it's not going to be that far off from anything that has to do with writing computer programs or doing anything technical like writing code - it seems like that thing could all be done by these algorithms very well and very efficiently. How engineers will use it as a tool or whether or not it will just end up replacing a lot of engineers is just up in the air. I think that's kind of the thing that's coming down the pipe probably." [Barry - H] "What I'm I guess more worried about and think more of on the short term is how AI is going to replace a lot of jobs. So you have the self-driving cars remove all truck drivers and taxi drivers, get them out of a job. Systems are also very good at searching and analyzing, well, human language as well, they're getting better at analyzing human language, which can actually hurt people like lawyers, which is almost surprising, as those I guess highly educated positions have never really before been threatened by technology. But maybe now they are." [James - U] "Trucks will become more autonomous in a way, and that's way more worrying for [the truck drivers], because we don't know what we are going to do with them." [Joe - H] "There's the broader future, which is somewhat scary. You know, AI, if it's sufficiently intelligent, will be smarter than any individual human. That's both fascinating but scary. …If an AI has been trained with the experience of a hundred thousand people, you know, it could unparalleled in terms of what it can create. I think it's already shown that it can not only do many tasks better than humans can, whether that be something skilled like a medical diagnosis, it can do those very well. But then people are like, 'Oh, it can't really touch the creative side.' Well, recent AIs have shown it can make its own pictures, sort of like a dream. It just draws from a bunch of different scenes and splices them together." [Justin - U]
Ambivalence	"[AI tools] model you so well that they can target you, they can predict exactly what is going on in your life. This is really, really scary, especially when it's not being used in good hands. …Actually, the technology helps	

	people, on one hand. On the other hand, it actually hurts people. And clearly, I don’t see a decreasing trend of this, it’s only increasing. …[If we achieve human-level AGI,] machines can replace humans, because if they can perform reasoning, that is one of the highest forms of intelligence. …And at that point, it becomes very mixed feelings, either you feel very thrilled about what you can do with AIs to solve all of these tasks, but also you are worried about what is the meaning of being a human at that point, because the machine is better than you.” [Bob - U] “For deep learning, it’s completely exciting, and it’s completely scary... I mean, the concern with general AI is going to be tough. So, what I say is that, yes it’s exciting and frightening at the same time.” [Zack - H] “The AI is going to change everything. Alexa, at home, Siri, in the phone. So, really, it’s the future. On the one hand, it was that part that, that’s, okay I’m going to get into the future. Now, I’m going to learn it and I’m going to learn it early. And on the other hand, I’m also a little bit afraid.” [Jeremy - H] “I think it should be positive. But there are always negative risks, right? …So if you have machines who understand the world better and in a similar way to how humans understand the world, that means that the machines can also understand humans better, which could be good because then they understand how humans look at the world. But that also might be that then they could do things to humans in the world or something like that.” [Sean - U] “Like the nuclear bomb, nuclear power has good things in terms of providing power. We discovered this first by science, and then the application was to make something that is terrible in some ways. This principle exists in nature. We just need to face it and try maybe to organize our society to deal with that better, but I don’t think that AI is such a threat.” [Henry - H]

	Unconstrained Creativity **(N = 18)**	**Heedful Creativity with Self-Imposed Constraints** **(N = 46)**
Relating to Outcomes	**Creating Psychological Distance from Negative Outcomes By:** **1. Anchoring on Present Moment (N = 13)**	**Commitment Manifested in:** **Excitement about Positive Outcomes** “I’ll always think a lot about theory and algorithm development, but the most exciting things for me I

	A. Technical Details of Work and Role Boundaries (N = 10) "I don't think it comes up on the day to day. I have a job, and my job is to make this thing work. So I'm so absorbed in the details that I'm not really thinking of the big picture. But then there's those times when you're just having some beers with friends, and you just get to philosophizing, then I like to think about these things, and they come up. But at the same time, if I'm reading stuff online, I'm less apt to read some pop-sci article about what some person thinks about the future of AI than I am to read about some very technical ML paper that introduces a new idea. Like that's what I'm just more interested in from a practical standpoint, is the details." [Justin] "Generally, I think it's actually not a technology or scientific question. But it's more about how the society and the policy makers can come to a consensus and then come up with a, you know, reasonable set of regulations. … So, all we really need to do is to ensure that there are going to be a group of people who are sufficiently fluent about this technology and then are not easily swayed by these new ideas – you know, papers uploaded on ArXiv every single day – or the people who have absolutely no idea. So, that's the group of people that we need to educate or, you know, to prepare. Not like, you know, the, 'Oh, okay, let's ask the scientists to wait a bit until we figure out this kind of thing, because we are not making progress that fast anyway.'" [Nathan] "I mostly work with how neural network learns, and if they learn badly, then I just retrain another one. The risk	think are on the applied side. To take these systems that work pretty well and use them to transform our understanding of what's happening in the world around us, or to make a positive impact with pretty small amounts of effort going into it, I think is really special and really exciting." [Doug] "There are really good potential use cases, like the drug discovery thing. It's like ,what are the important problems for the human race? Take out profits, which is like, being at a company, you cannot say that, right? But I don't care about how much money we're going to make off of this, do it because it's the thing that the human race should be doing that we have to do, or we go extinct. Like, there's all kinds of problems that we should be exploring as fast as possible. How fast can we get machine learning to help us cure cancer? How fast can we get it to eliminate poverty? Or fix global warming? Like, there's so many problems to solve in the world. …I think, like, medical stuff is the best example because that could eliminate so much suffering in the world." [Barry] "I think it may be that we need an open-ended process to actually produce something like a brain. …I'm pretty confident we can get really good, open-ended process within about 30 years. And that would be amazing. You could imagine artificial systems proliferating with cool, interesting stuff forever, like on the level of what earth can do, this would be an amazing step that would be very interesting and maybe help us get to AI." [Ryan]

	associated with my work is reasonably small. So maybe that’s why I don’t worry about it.” [Jenny] “I mean maybe it’s true AI could have this negative effect [if segments of workforce are displaced]. Maybe it’s just something that we will have to take into consideration. It’s not something you can stop. It’s something you can just think about and maybe… You know, many things in our common life, in our world, are modeled after a very old idea of life and work, like the 8-hours-per-day work in an office or a factory. Those were just decided arbitrarily during the industrial revolution. And maybe they make no sense now. So maybe that should just be changed according to the new progress. But that’s for someone else to decide. … So I trust the scientific community – this is what I always do – to come up with good guidelines and solutions for this.” [Nick] “I feel that most of our technology is probably going to be highly protected, at least if we’re smart about it. We’re going to keep building systems for particular purposes.” [James] **B. *And/or* Limitations of Current Technology (N = 6)** “We made this huge program, and we had these numbers that seemed to go higher than humans, but on the other side, you had these large weaknesses, these weird things about the system that I guess we don’t understand that well. These systems are not as powerful as maybe they seem at first glance. So I guess that is why I think this is going to be more gradual. …I feel it’s not going to be this sudden, suddenly AI takes off and takes over the world. I	“Something that many people in the field might say is that they want to make a better world. A lot of people say that. But that’s too vague for me. So, but it’s in part true. …I want to build things that benefit ourselves, that will improve the lives of people.” [Jeremy] “We often think about unreal spaces and solving problems in those unreal spaces. Or transforming problems into unreal spaces, which we search. …And then you slowly realize, hang on a sec, this thing is actually really quite good, and look, it can solve a lot of these standard data sets, and if it can solve them, maybe it can solve some real stuff, and, oh yeah, it really can. So personally I’m a bit more practical and I like to think about the impact and I like to think about what problems we can solve.” [John] ***And* Concretizing Negative Outcomes** “Humans seem to be able to deal with edge cases. You know it’s like if a horse flew out of the back of a car in front of you or something, you would still probably try to get out of the way. But a computer might not know what it’s looking at and just do something crazy. It could do something crazy.” [Ryan] “Someone… asked [a journalist] a question about automation of labor, and she said, ‘The reason we’re here now is not because automation is coming. Automation has been affecting people’s jobs for the

	think there’s a fairly gradual process with a lot of I guess intermediate episodes where we can get some additional experience with what AGI might look like and get more informed of how we need to deal with it.” [James] “Humans are actually very, very good at being flexible and recognizing rules and figuring out when things apply and when they don’t. And to the extent that we can codify computers to mimic that, it’s kind of getting better. But that’s a lot of the reason why I think AGI is so ridiculous, it’s because it is very much on this level of, we train something to predict, you know, 60 numbers. And this is a far way out from that, and even knowing what AGI looks like, cause we’re still not at that point. …Generally the further we’ve dug in this space, the more complicated things have become.” [Charles] “Silicon Valley’s very exciting. It’s a very exciting place and a very optimistic place. And I think I tend to be a little bit more on the I guess skeptical side. I’m aware that like as someone who’s down in the weeds, I might be overly obsessed with seeing all the difficulties along the way instead of thinking about the broad picture.” [Andrew] “We often make a lot of backwards steps without realizing it and, you know, after about several years, we realize that oh, actually, we took a wrong path, you know. We’ve got to go back for something. So, to me, it’s actually not that fast. I don’t worry about it. We are making a lot of mistakes anyway in science. So we are not moving that fast.” [Nathan] **2. *And/or* Emphasizing Positive Outcomes (N = 8)**	last century. The reason we’re here now is because you’re worried it’s going to affect people like you and me.’ …And it really struck home with me. And I began to think more – you know there’s a tendency in my field, and I’m particularly guilty of it, to sort of trivialize what we do. So, you know, I work in video games, and I don’t even work in particularly applicable research in video games. So for a long time, I just thought, ‘It doesn’t really matter what I do; no one’s ever going to pay attention. Some science journalists will write about it; that’ll be nice.’ It began to be clear to me that … actually my work was not as trivial as I thought, and there were real applications.” [Alan] “The first, which I think might be a problem, it’s about employment. I always think about it with the truck drivers. …AI will help for economic reasons. …But then the question is, what people are going to do, who will lose their job, what are they going to do? And I think everyone has heard of that, from let’s say, AI experts. And I have to say I think it’s a real issue. I have no actual answer to this issue. …And we have to think about it. …I’m developing, for example, if you take the truck example, sometimes when I work on this connected truck project, I think, yeah, one day, maybe these people won’t have jobs anymore.” [Joe] “If it’s a classification problem, people are going to try to use this stuff for it at some point in the near future. There’s some classification problems where it’s, like, not necessarily a good thing that it gets used, because it removes the human element. Like, how I

	"I think it's really promising, the things – I mean again it's scientists who try to use deep learning for different tasks. And so that's one thing that I think will be progressing quite fast in the future. And also … that the networks can do everything that we cannot for them by themselves – so they can adjust their internal architecture and settings by themselves, I think this will also be a promising thing in the future." [Andrea] "We want to create things that does things we're not good at so we can concentrate on what we're good at. So I'm quite optimistic." [Mark] "I'm pretty sure we will have more cobots – so robots which are working together with the human – for long time. And the robots which will replace the human are already there. And for me, for sure, we will have more and more and more robots, etc." [Brad] "When this new generation of kids grow up, it's more embedded to their daily life right now. So almost everyone grows up with technology. They almost aren't aware of it, because it is embedded into their life. So I think – I guess it will co-evolve with the technology as we use it more and more. …I still believe I've learned AI will be very important in our daily lives for a lot of small tasks." [Eve] "Garry Kasparov has this thing that he said where like a computer can always beat a human, but like a human and a computer working together will always trounce a single computer. And I think it's a really interesting idea to find ways to use AI alongside humans to empower them. So	got into that question is, my sister is a social worker, and they have this recommendation system … and it spits out a list, a rank-ordered list of who the homeless people are that need housing the fastest, who the highest-priority people should be. And they're required to use that list as the list, to fill it in that order, put the person in number 1 in the first house that's available, put person number 2 in the second house that's available. Those are kind of other scary uses for it because it removes responsibility from the person that has to make that decision and instead of going out and interviewing the person, they're just looking online at text." [Barry] "…You have other rogue elements I guess you could say, like not in the US Military but somewhere else, some other military, like a non-country type of entity, where they would just want the AI, they don't care about whether they're ethical or not. And so those kind of things, that's a little tricky there. So I guess the hope is that you are not going to pass the technology around in the public if there is such a technology. Like, if there's technology that could wage war, like, automatic – autonomously, that's just not the kind of tech that should be open-sourced on the internet. But then, you know, it's not completely clear when you have created such a technology. Like, if you just have generic AI that's super intelligent, if that's in the public domain, people could use it for whatever they want. And how do we prevent it from being abused?" [Ryan]

	for instance instead of the AI writing songs, there's so many ways that an AI could help you write a song. And it sort of combines human creativity and AI creativity and I think that could be amazing." [Jacob]	"So when you no longer have to send soldiers into harm's way to wage warfare, when you can just send your robots or whatever, or just use your robotic drones to make the air strikes, whatever the particular technology might be, that makes it that much easier to do completely horrible things to each other. So that is certainly a big concern from my standpoint." [Ellen]
Amplification Dynamics	**Amplification of Positive Orientation** "We will have automatic art generation. …And this is very interesting that in the future we will have a new form of art that will be generated by AI, and that puts a lot of good creation and interesting creation on the table. I'm very enthusiastic about this. I think it's very good, because this is new tools for artists. So of course individuals will use these tools to make art without artists, but I mean that wouldn't be very interesting art. …What is very interesting is this new electronic music, new art, and all the questions that are raised, I think artists will have new challenges, and can use these tools as new content to explore. And I find this very nice. Inspiring." [Dennis] "If a recommendation is for something you may not already know and how to find that information and make it available to each person, I think that one part excites me a lot. And the other thing is if you think about the current recommendation results in your daily life, it's still far from satisfactory. There is still a lot of room for improvement. So that's another thing that excites me." [Eve] "I've stayed optimistic. I think technological development is always a good thing. It has bad effects, it generates bad	**Temporary Strengthening of Negative Orientation: Urgency and Responsibility for Negative Outcomes** "If we're just responsible with it, and it comes down to the question… These language models are gonna pick up on a lot of bias that we write, now, as humans. …If your entire existence and all you were and everything you ever knew was only the things written on the Internet, maybe if you think about it like that, yeah, maybe these models are gonna be amplifying a lot of the things that we already don't like about certain things we do. And that's a real problem. …Like, how do we, just as responsibly, how do we be careful with the data?" [Harry] "There are things that we can do, but maybe we should just avoid. I'm just trying to think about it in games specifically. There are a lot of people who think that games are a good way to teach, like, political ideas or solve political problems, which I'm very skeptical and worried about and avoid generally. …I'm slightly worried about what the implications of that might be. …It's responsibility; a big part of it is. …If I also can talk about social issues and labor issues and political issues, then I think … dispels the

	outcomes just as anything – if any movement is not handled well, it would make some aspects of society horrible, but I think in general it’s moving towards the right trend.” [Jenny] “When it comes to for example content generation, when it comes to job replacement, I’ve thinking in terms of, well, we have this amazing method that I’ve been helping with for generating income for example. Will this make designers unemployed? ‘No’ is my answer to this. …Actually, we’re creating new tools.” [Mark] **Constraints Unnecessary or Counterproductive** “[My concern about dangerous applications is] more in the back of my mind. If I was letting it influence things, I might not be working in this field. I might be afraid of it and not trying to push it. And in fact I’ve got a project in the works to make these generative models much better than they are currently today. So I’ll probably be part of the problem.” [Jack] “Maybe at some point we come to a point where there simply isn’t enough work for people to do. So at that point basically becoming unemployed is just – it’s not the fault of the person being unemployed, it’s just like, there’s not anything to do because all the work has been taken over by AI. And I think something like that will require a huge shift in society in general. …And if that starts happening to more and more people, and no more jobs get created, we will need to adapt as a society. …It’s something in the back of my mind, but it doesn’t affect my day to day	idea that you can just do AI without engaging with these things. There is no such thing as an AI account with just tweets about AI. You have to do all the other stuff because that is the other half of AI.” [Alan] “The things that people tend to think of immediately is what are the potential harmful effects of AI on humans, but they never want to go look at the other side of it, which is what are our responsibilities to those AIs? I mean if these are really human-level intelligence, what does that actually mean? Is it ethical, if they get difficult, to just pull the plug? Is that ethical? …Well if they have human-level intelligence, are they just machines anymore? So this is very much the part of the ethical thinking where I am right now. This is what has really driven me away from the notion of human-level intelligence. Or for that matter even cat- or dog-level intelligence. It’s… where are those lines? …Just the whole idea of creating these AGIs, but keeping them limited, keeping them under control, is like, so it’s ethical to create a whole race of slaves? That’s an okay thing? I don’t know. I don’t feel particularly good about that. …We do get very tied up in the ‘Can we do this?’ and don’t take the time to step back and reflect on whether we should. Because they’re too very different questions. AI is an area where if we fail to examine the ethical foundations and ramifications of what we are doing, I think we fail not only at our own peril, but at everyone’s peril. They really are critical questions.” [Ellen]

	work, at least not at this point. …It's even hard to see what should be done, if anything." [James] "And most of these far-out [news] pieces, honestly from my perspective, I think they're kind of ridiculous. I understand that Elon Musk obviously gains a lot from convincing people that [AGI] is a realistic thing to be worried about. But, it's not my view on it at all." [Charles] "I really believe that trying not to develop something is useless. Because I mean, that's really, I'm not that unique, my capabilities are not that much greater than anyone else's. If I don't do it, others will." [Mark] "How we do research is to come up with methods for generating something new and also to come up with a way to distinguish whether the things that are generated are things are good or bad and so on. Essentially, this kind of research or science is almost always, it's a double-sided sword, right? So, you know, but the thing is that it's impossible to make a sword to have only one side – that is sharp on one side and dull on the other side. It'll always be, say, sharp on both sides. So, generally, I think it's actually not a technology or scientific question. …The applications and how these things are going to be used are really, really, fun to think about, obviously. But the things that, how these are going to be used is, in my opinion, is not really up to me or to my imagination." [Nathan]	"There's a lot of danger in placing too much faith in imperfect algorithms that are, that… I think it's one thing to recognize that they're always imperfect. The bad thing is when you assume – when you act on it knowing that it's imperfect and that affects a lot of people's lives. So bad credit scores and ratings systems, there's all sorts of problems with encoding biases into ML - encoding the biases that people already have in an automated way rather than trying to fix those biases or make the automated version better than the humans. There's a good book on that, "Weapons of Mass Destruction," that details a lot of ways that algorithms can be misused to ruin people's lives. And that's a huge danger. If we place too much faith in single algorithms that are too centralized, then it's not a great thing for society." [Colton] "I guess [I have] maybe a sense of responsibility I think. So I am in a position where maybe I know all these beautiful techniques and I know at least for some of them that they could be used to do wonderful things. And so maybe I should push a bit more in that direction to help." [Dylan]

Categories and Types of Self-Imposed Constraints

Positive Moral Valence (N = 25)	**Constraining Project Choice** “As an AI researcher, my own thinking about what our goals actually should be has actually changed quite a lot in the past 10 or 15 years, since I really started studying the field in depth. I guess as a field, for AI, the holy grail, if you will, is human-level intelligence. …As a researcher, I deeply question whether that should be our goal at all at this point. And that’s purely from my own ethical standpoint. The way I think about it is, we can’t treat other human beings particularly well and equitably yet across the board as a species, so how would we treat an intelligence that was completely different? So I’m really questioning that whole idea of human-level intelligence, if that’s really what we should be after at this point.” [Ellen] “The way I typically choose the applied field is looking for something that I think is a positive contribution to society. …Certainly the choices to pick applied projects that are inherently and undoubtedly good for society, I hope that I would have done already, but certainly having a constant reminder of the pros-cons list I think definitely helps nudge things in that direction.” [Doug] “I don’t want to make AI that is, for instance, AI related to military stuff. So, the thing is that, I signed … a letter that many or most of the main AI researchers signed, I also signed, about asking to limit the use of the AI in the meantime.” [Jeremy] “Leaving (eliminating) the human factor from the war is very dangerous. So I don’t think it’s a good idea to develop autonomous weapons. On the short run it might be a win, but on the long run it’s certainly a big issue on how humanity might develop.” [Christian] “[A colleague and I] were talking about AI that would be able to play games and give feedback on them. And she said, ‘You know, if an AI played a game about grief or trauma, any kind of feedback it gives on it is kind of going to be weird.’ So there are a lot of issues about, are there just hard limits to what AI can do? Is it – even in our best case with all of these considerations and being as understanding as we are trying to be, are there actually things that society will never want AI to do? And I think about that question a lot. Or are there things that AI should never do; that’s another thing. There are things that we can do, but maybe we should just avoid. I’m just trying to think about it in games specifically.” [Alan]	
Control (N = 21)	**Constraining Functionality (N = 18)**	**Requiring Full Understanding (N = 6)**

	“[We’re working on] deciding which is the right action to take in this case of extreme danger or extreme rare occasions. You’re doing something in which a machine has to select killing one or killing the other one. …In the next months, I guess, [we’ll be] at the point where we’re going to insert this knowledge into the model.” [Jeremy] “There’s kind of this lab-experiment environment, where I just go and I spend some time just trying to make the model as good as possible, and I don’t think about those things. And then, there’s this compartmentalized time, where we step back, and we say, ‘Okay. Now that this model’s really good at the task we want’ – so, let’s say, language generation, ‘Now, how do we almost make sure that it doesn’t generate things we wouldn’t want it to generate?” [Harry] “I think a human can also react. It’s not just AI who has action on this. So a compromise could be, if we are not sure, we will ask some human to do the operation, to take action. …I think it is very important for people to have the possibility of being involved in the AI system. Even in some kind of use case, it is not so natural, but I think in nearly every AI application, humans can be involved. For example, in fraud detection system, the system will give back fraudulent score, ranking. But the system will not do any decision. It could be still the human who does the decision and takes into account the feedback given by the detection system. But the final decision will be given by the human. So still I think it is important to make sure or to make possible that if we want the operation involving of the people, we should give the possibility. I think for the moment, there are no	“One short-term, proximal sort of thing is, can we just get more insight into what our current agents are actually doing, what they’re learning? So it’s kind of like a safety issue, but it could be a basic scientific issue as well. …It’s a recent sort of neglected area of research.” [Adam] “I’m primarily focused on making machine learning models interpretable in different ways. So a lot of the time when you’re developing an ML model, you’re kind of just looking at the accuracy or certain statistics about it which are informing you in some way of how good the model is. And that’s sort of how you develop your – what works and what doesn’t. ‘Okay so the number went up that I know means this, that means that the model’s good.’ But that can lead to a lot of problems. So my work’s basically about making the internals of the model more clear and kind of – instead of there being 20 million weights or 20,000 weights, each of them associated with a number which has some importance, there would be you know clear concepts, and then those concepts would be increasing or decreasing in a certain value. …So and basically filling in the middle step which explains what the model is doing. …I’d say I don’t fully understand everything that I create. Because I guess with linear models, so these are models that you know you can see every single step of what they do and you can see how they’ve done it, so it’s like a decision tree for what I was saying before… That’s quite easy to understand. I can see directly the process it’s taken to get that output. But any other model, where it’s just input – don’t understand the middle bit where the model works it out – output, I’m always in a state of, ‘I know this, maybe this concept, which is a little abstract, and maybe I know my data quite well, I know what kind of data it is, I know how it’s represented,

	applications which can achieve full precision, 100% accuracy [without humans].” [Marie] “We believe that the human in the loop is a more perfect system than the machine – at this point, right? Maybe in the future where AI evolves, and you have general AI that is very powerful, then you could have that. But I think there will be a lot of time before that happens, because people I think want to still be involved and double check. There’s a trust issue.” [Evan] “Personally, I do think that you have to strike a balance. …Because of its dual-use nature, there are a lot of bad uses for AI technology. So, I actually wrote some papers about robustness of neural networks, for instance, and how you can try to safeguard yourself against sort of adversarial attacks. And I think in general it is an area that we’re very cognizant about, and I think that’s extremely important that we understand the controllability of the models that we develop, and how we might control or put [safeguards].” [Caleb]	and I know my outputs – how my outputs were labeled, kind of what they are, but this whole middle bit, I don’t understand almost anything about what it’s really doing.’ Or at least it feels that way.” [Gary] “So we actually try and be really honest about [Game Design AI]’s weaknesses and the system’s shortcomings, and we’re very open with its process. Like, I literally show people how it plays games; you can just watch it. And the intention there is for people to realize that actually it’s not intelligent in the way they think it is and then have an appreciation for the ways that it is intelligent.” [Alan] “There’s a lot of issues right now with machine learning being a black box up and down the expertise chain. Like, Google doesn't know what's going on in some of their AI. Every big AI company is trying to create tools to help them figure out what the heck is going on. …AI doesn’t have to be a black box. Some of the machine learning stuff, at the moment where our technology is, is a black box because we haven’t figured out good ways to look at it yet. Nothing is hidden. You can always look at all the ways. …So it’s not a black box, it’s an open book.” [Joy] “A lot of this work into interpretability and explainability, most of the time, when you force a bottleneck in the model somewhere, where it has to do something like this, the overall performance of the actual task tends to get worse. But, in this case, we found that it actually improved. So, the way that we did it had some benefits to saying, ‘Hey, guys, everybody else in the community, guys and gals, if you want explainability and interpretability, maybe there’s a way forward, without losing on performance,’ which is what

			lots of people care about. And that’s kind of the first goal that a lot of people think about. And we were showing, ‘Hey, there’s a way that we can kind of put something in the model where it kind of looks like we’re seeing what it’s thinking, a little bit.’” [Harry]
Quality (N = 23)	**Constraining Input (N = 10)** “We need to be careful about what sort of data we provide to AI, how biased the data is with regards to gender, to culture, to race, to all these sort of biases that we tend to – unfortunately, we tend to use daily, as humans. So, we tend to actually not overlook these biases, but when we provide it to AI, AI’s nothing more than sophisticated statistical system, especially machine learning, the machine-learning aspect of it. Sophisticated statistical learning system that will learn from the data you provide. So, if we provide garbage, AI will learn how to behave in a garbage way. So, we need to be careful about that.” [Steven] “It’s much more important to understand the type of data they have or could get – these two things are much more important than which specific machine learning technology or AI we are going to… I mean once we have the data and are confident …[that] the data actually representative and so on … it’s something we’re very aware of.” [Nelson]	**Constraining Own Reactions (N = 6)** “When you’re making a prediction, and the prediction turns out to be true, it’s very important that still you could say there are hidden information, and you have to doubt everything you said and did and do it again, to really be sure that it’s true. And, even with that, be cautious about this… Do not be over-enthusiastic.” [Zack] “I think we should be a lot scared by black boxes that we think are intelligent. It’s very dangerous, because we want to project something about this being intelligent, whereas most of the case it’s not. And if we assume it’s a black box that’s as intelligent as a	**Constraining Release Process (N = 10)** “…We should keep our heads and remember, okay, we still have to do testing, we still have to do certification, we have to do retraining, we have to look at all the implications. You know, we know how to do this stuff. …We have a lot of safety procedures, some of which may seem tedious and boring to the general public, but actually they’re there for a good reason, because they do keep us alive. And therefore if we’re inventing new technologies, which will affect our safety, whether we call them AI or not, then hand-in-hand with that, we have to say, okay, what sort of testing? What sort of safety certification should we be having? How do we guarantee that we can actually rely on these technologies enough to use them? And it’s easy to pick on the self-driving cars, because they make the headlines. But there are unfortunately very readily available examples where people have believed the hype, and they’re sitting behind the steering wheel reading a book, and their vehicle has unfortunately been proven not

	"These systems act on knowledge. I can think of scenarios where, you know, if they do get access to knowledge that's either sensitive or get some high level insight that could be very dangerous to have, then these kind of technologies can either be – if they're concentrated in one place, that can result in, you know, differences between nations and the world or things like that." [Noah] "At one point, our team was working on this model that would predict jobs. You will see recommendations for jobs and so on. So at one point, there's an issue with diversity. We said, 'Look, we have these features that are potentially discriminatory, or people might find them objectionable.' So that was quite emotional. Should we have age as a feature? Should we have gender? Things like that. That elicited a lot of strong emotions from people – it's quite personal. These concerns were addressed, and we took steps trying to make sure that our models are as unbiased as they can be. There's always going to be bias, but we're trying to make sure that there's not something that people could object to." [Greg] "So one problem you can have in these very class-imbalanced situations, where it's much, much more common that you don't have cancer than you do have cancer, is that the model will default to predicting the most	human, I'll assume a few things that will be completely wrong, and that will be very dangerous. So we start putting these kinds of black boxes everywhere. And they are all biased, they all have bugs, they all have issues. And they are much less smart than we think. And that's one thing that will be very dangerous in the future, is when we start relying on these things without having an understanding of what they're actually doing." [Daniel] "…It kind of scares me because of the implicit trust that gets put into these. …We're only starting to understand the negative impacts they could have with that implicit trust in them." [Kevin]	to be safe enough and has got itself into trouble through whatever environment wasn't anticipated, and the person who should've been a driver but in fact was a passenger goes and gets themselves killed. We don't want that kind of situation." [John] "It's become more and more… I don't get results and then send them to my supervisor. I get results, and then I double-check, and then I triple-check, and I make sure that all of it clicks together the right way. And then I do a proper debugging process, so I have a bunch of tests that work. I know that I expect a certain result, and I expect them to behave a certain way. So that, given an input, I know that if that works, then maybe my model has a higher chance of working. So it's just an iterative process of finding the problem and then solving it in some way and testing it in many different ways so that I'm sure this is the right thing." [Gary] "When you train a machine learning model, …we decide whether or not we're going to release this machine learning model to production. There's this process of evaluating what's good enough. So is this model of sufficient quality to release into the wild? And this is particularly

	common class, so in other words the model can achieve a very good error rate simply by saying ‘You don’t have cancer.’ Although this might be optimal for optimizing the error metric you’ve defined, it’s definitely not optimal from a human point of view. So you may want to change the input data you provide to the model or change the way the model works, such that it produces results that optimize for a different metric, or are closer to what you’d want for future predictions. So there are kind of these failure modes that you kind of have to avoid, that aren’t immediately obvious, I guess. What I’m trying to get at is, you need to very careful about which error metrics you’re optimizing for, so there’s a translation task there that happens between user and machine.” [Richard] “There’s always a question coming up, ‘Well, what are the biases in the data you’re using? Like, how do we identify those? How could you remove those?’ …How do we almost make sure that it doesn’t… amplify bias that is kind of found in data that we collect from the wild?” [Harry]		important for tasks that are very subjective – so tasks for which there is not necessarily a correct answer, or the correctness of an answer is a bit more unclear.” [Richard] “Things are slowly progressing. What I’m doing now, basically I’m piloting the studies on the young people that are available in our group. So what we are trying now, we are trying to apply our AI techniques on the people over here. And then I think when it’s ready, or when we are confident enough, if we can get some data of the elderly people, then we can also apply on them. So I think it’s better to pilot things on ourselves first.” [George] “So as people will still go very slowly, at the same time we can have more time to better develop the applications. Also this will give more time for government and actors to propose more regularization, to make sure that AI systems will be very secure, will be very performant.” [Marie] “I’ve been involved in some testing of technology that have been involved in European Space Agency and Mars Rover, right? Ok, when you’ve sent it, you’ve sent it [laughing], right? You then should spend more time and use a lot of

			resources ensuring it's more correct. But it's still not going to be perfect." [Nelson]

	Unconstrained Creativity **(N = 18)**	**Heedful Creativity with Self-Imposed Constraints** **(N = 46)**
Engagement in Creative Process	**Engagement Facilitated by Positive Orientation Remaining Amplified** "A lot of this uncertainty is a core part of why I like this space. There's maybe some reasons to that, and maybe part of it is kind of fun. …When you jump in at the deep end of the pool, it's really not so hard. So it's a really nice place to play in." [Jack] "Some of them are just beautiful. So algorithms help solve problems faster. …They seem like art, like really beautiful to me. And intriguing in a way you have no idea how they were formed in the beginning from some normal people's mind. …We now are doing [our work] in a way of, 'We don't know how it works, let's just test different things and see how it responds.' But that's how most natural science is done, so I don't see a big problem there." [Jenny] "When there's no science behind it, so you do not know what the outcome will be, then it could be good or bad surprise. …With deep learning it's more like a surprise. Which I think can be good and bad, but I like to see	**Engagement Facilitated by Positive Orientation Strengthening** "I'd say I'm positively surprised almost every time I look at the explanations that we're able to get from a model, or at least when they do well. If I'm looking at an interpretation of a model, and it says some interesting way of looking at how to get a genre of a movie or a sentiment of something, then I'm surprised. I'd say, 'Oh wow, the model has learned something really interesting, which is enjoyable to read.' I suppose the times I've been positively surprised are when I've actually looked at how the model is doing something, and it's doing it really well in a new and interesting way that I hadn't thought about before." [Gary] "Gosh, what are the feelings you get when you do that? For me there's definitely some of that compulsion, just, like, the need to see what's next. Sort of almost like … Advent calendars or anything where there's a surprise as to what comes next that keeps you hooked. …And I find a lot of the fun of it is curating what I'm putting into my things, like grammars. So I know what's in my generator, but then it's kind of an interesting automated juxtaposition reflected back at you when you when you see stuff that's generated. …If you put two things together, you get something bigger out of that. It's kind of the emergent magic of a lot of, at least my work, where there's a lot of juxtapositions." [Joy] "I think it also is, like, really fun to see what the algorithm will say and the mistakes it makes. Not even thinking, 'Oh why is it making those mistakes?' But …I think a lot of people will attribute a

	the good part, it can be interesting, challenging, promising.” [Andrea] “I think that’s actually one of the most fun things about doing research in AI, is the black box nature – that sometimes you get solutions that, yeah, you didn’t expect. What are *those*?” [James] “It’s like a playful science experiment, and you record your observations as you might be like kind of studying new phenomenon or a new creature, not really knowing how exactly it works.” [Christopher] “Just to understand intelligence, what is consciousness? Is it possible to create artificial consciousness? …And finally, the last motivation is to have fun, simply because this is very fun, to implement machine learning algorithms to work on robots. This is like playing with big toys.” [Dennis]	personality, or some level of like, ‘Oh that’s kind of funny that it would do that,’ and we’re talking about it less like an algorithm, more like a creation that we’re thinking about. So there’s some life to it, and that’s why I’ve always kind of liked machine learning.” [Christopher] “When you’re actually producing sentences or producing images that look life-like, and you’re kind of digging into the imagination, quote-unquote, of a machine learning model, that I think has a huge emotional reaction. … That’s what starts to feel like magic, that’s what starts to feel like creativity – having a model produce something perhaps no other human has said before but just sounds right and feels human. …It’s quite honestly mind-blowing to me that you can kind of define a very simple set of rules and hit go and watch as an algorithm picks up on patterns that you wouldn’t necessarily see yourself.” [Richard] “Working through ML problems and working with AI algorithms is exciting and interesting also, outside of these moments of frustration and joy. That’s why I like to work with this, because it gives you lots of excitement along the way… I guess the whole process, from like learning about a new idea – hearing and being inspired by other people’s work and their techniques in ML, to coming up with your own ideas and own hypotheses and testing those out and analyzing. It’s a creative analysis to try to understand models that you build. So I guess the whole work process, from learning new stuff to creating ideas and to analyzing and interpreting results.” [Kenneth] “The deeper the abstractions you can learn, is kind of how I frame it in my own head at least, the cooler it is, roughly. I think deep learning is really, really fascinating. I think the more unsupervised the process for kind of like getting a structure of a model, the more interesting of an AI primitive you’re working with. And I believe

		that on a pretty base level. I just find the deeper ones cooler. …It was just like, kind of wonder at the coolness of the thing.” [Frank]