**From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Economics

2018

# Essays in Behavioral Economics

Young, Benjamin Garry

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp01q237hv648

*U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf*

# Essays In Behavioral Economics

Benjamin Garry Young

A Dissertation

Presented to the Faculty

of Princeton University

in Candidacy for the Degree

of Doctor of Philosophy

Recommended for Acceptance

by the Department of

Economics

Adviser: Professor Roland Bénabou

June 2018

© Copyright by Benjamin Garry Young, 2018.

All rights reserved.

# Abstract

This thesis uses economic theory to investigate two important behavioral phenomena: the fact that the beliefs of individuals can be distorted from the truth, and that individuals can endogenously transition between different modes of cognition which affect the rationality of their decision-making. These ideas are looked at in three distinct applications. Chapter 2 looks at a setting in which a contract designer can distort the beliefs of an individual away from the truth through how he presents or frames a contract he is providing. The susceptibility of the agent to these belief distortions is a function of her mode of cognition, which is endogenously determined as a function of the frame. The model generates novel predictions due to the equilibrium effects resulting from the joint determination of framing and cognition. In particular, it is shown that cognitive investment may actually be increasing in its own cost. Chapter 3 investigates an intra-personal setting in which an agent can distort her own beliefs through the use of different types of self-set goals, and may or may not find it advantageous to do so. This is used to provide a foundation for why individuals repeatedly set goals for themselves that they deviate from. It is shown that individuals with large self-control problems will utilize goals that distort beliefs and, thus, will end up deviating from their self-set goals. Finally, in Chapter 4, the theory of endogenous modes of cognition is employed in a communication setting. It is shown that bounded cognition can actually increase the propensity for information to be transmitted in equilibrium. A paradox of cognition is established: the greater the cognitive ability of the agent, the less information is revealed to her in equilibrium.

# Acknowledgements

I would like to acknowledge the numerous people that have supported me over the course of my Ph.D. at Princeton.

First, I would like to offer my deepest gratitude to my adviser, Professor Roland Bénabou, who has offered invaluable guidance over the past six years. My passion for behavioral economics was awakened when I took your graduate class in my second year. Before this, I had little working knowledge of this fascinating discipline of economics and it was your instruction that inspired me to alter my academic career path to work on these topics. Having you as my adviser, being able to work so closely with you, and understanding your approach to economic thinking has made me a much stronger economist. Words can not express my appreciation for the amount of time you have dedicated to my academic development and your mentorship will be an important component of any success I achieve in my future career.

I would also like to acknowledge the remaining members of my thesis advisory committee. I would like to thank Professor Stephen Morris for serving as a reader of my thesis and for providing me with the opportunity to work closely together over the years in a teaching capacity. Professor Wolfgang Pesendorfer and Professor Pietro Ortoleva both offered useful feedback on my research in the lead up to the job market and I appreciate the time you gave me in this regard. Finally, I would like to offer a special thank you to Professor Leeat Yariv. You were instrumental in the job market process and I could never express enough gratitude for your support. I wish that I had the chance to learn from you and work more closely with you during my time at Princeton. In general, I will miss the support and vibrant discussions of the brilliant theory group at Princeton.

I would like to thank Princeton for its support in the form of funding and resources over the past six years. Princeton has made life as a graduate student as easy as it can possibly be and this fosters a fantastic research environment. I would also like to thank Dr. Hamid Biglari for funding a Behavioral Science Fellowship which supported the fourth year of my Ph.D. This was extremely valuable for ensuring the quality of my research was as high as possible. Finally, I would like to thank my department's graduate administrator, Laura Hedden, for all of the encouragement she has given to all of us over the course of our graduate studies.

I would like to acknowledge the undergraduate students at Princeton for making my time as a teacher so meaningful. Being in the classroom has been one of the most enjoyable experiences and this is due to your insatiable appetite to learn. I would also like to thank Professor Dilip Abreu for providing me with opportunity to be heavily involved in the teaching process, as well as the advice you offered over the years. Your presence is sorely missed at Princeton.

Many thanks go out to those that I have met and become close to over the course of my time at Princeton; you have all made moving from Australia to the USA so easy. Most importantly, I want to thank Martina for being a solid foundation during all the ups and downs of the Ph.D. Your approach to life has been an inspiration and I would not have been able to survive this process without your love and support. I am so happy to have met you and you have made me a stronger person. It has also been a pleasure sharing this experience with all of my fellow graduate students in the economics department. To Emil and Graham: I am so happy that we have developed such strong friendships and even though we will be miles apart from here on out, friendships like this last a lifetime. I want to thank Alex and Justin for putting up with long discussions about my research or just anything that was on my mind. And

to Luca; I am grateful for the circumstances that made us friends and I was honored to be the best man at your wedding.

I would like to acknowledge those that pushed me to undertake the Ph.D.; Professor Andrew McLennan, Professor Shino Takayama, Professor Flavio Menezes, and Professor Fabrizio Carmignani. Without you I may never have found this career path and I am extremely grateful for that.

Finally, I would like to thank all of my family and friends back in Australia. Even though we are separated by thousands of miles, I have never truly felt like you were that far away. I want to express my eternal gratitude to my mother. You have sacrificed so much for me to ensure that I could pave whatever path I wished in my life. I could never have been in this position without this unconditional love. I wish to express a similar sentiment to my grandparents. I feel so fortunate to have as close a relationship with you both as I do and appreciate everything you have done for me. Finally, I want to thank my amazing network of friends, Ben, Rory, James, and Byron: you have always been there for me and while I wish we were able to talk and see each other more, I know that our relationships will always survive distance.

# Contents


Abstract . . . . . iii

Acknowledgements . . . . . iv

List of Figures . . . . . xi

**1 Introduction** **1**

**2 Framing and Cognition in Contract Design** **6**

2.1 Introduction . . . . . 6

2.2 Related Literature . . . . . 11

2.3 Model . . . . . 15

2.3.1 Primitives . . . . . 15

2.3.2 Discussion of Model's Assumptions . . . . . 20

2.4 Single-Contract Offered . . . . . 24

2.4.1 $f$-Frame Equilibria . . . . . 25

2.4.2 Optimal Frame . . . . . 34

2.5 Screening . . . . . 41

2.6 Conclusion . . . . . 44

**3 Goal-Setting with Endogenous Awareness** **47**

3.1 Introduction . . . . . 47

3.2 Related Literature . . . . . 53

3.3 The Model . . . . . 56

3.3.1 Formal Details . . . 56
3.3.2 Discussion of Assumptions . . . 61
3.4 Case of a Binary State-Space . . . 64
3.4.1 Complete Goals . . . 64
3.4.2 Incomplete Goals . . . 70
3.4.3 Optimal Goals . . . 76
3.4.4 Goal Deviation and Welfare . . . 79
3.4.5 Further Comparative Statics Results . . . 84
3.5 Arbitrary, Finite State-Spaces . . . 87
3.5.1 Optimal Goals for a Highly Time-Inconsistent Individual . . . 87
3.5.2 Optimal Goals for a Fairly Time-Consistent Individual . . . 88
3.6 Discussion . . . 90
3.6.1 Allowing for Goal-Abstention . . . 90
3.6.2 No Psychological Disutility to Self-0 . . . 92
3.7 Conclusion . . . 93

**4 Communication with Endogenously Naive Receivers 97**

4.1 Introduction . . . 97
4.2 Related Literature . . . 99
4.3 The Model . . . 100
4.4 Main Results . . . 104
4.4.1 Cognitive Best-Response . . . 104
4.4.2 Period $t = 2$ Equilibrium . . . 107
4.4.3 Period $t = 1$ Equilibrium . . . 108
4.5 Conclusion . . . 113

**A Proof of Results 115**

A.1 Proof of Results in Chapter 2 . . . 116

A.1.1 Proof of Lemma 2.1 . . . 116
A.1.2 Proof of Lemma 2.2 . . . 118
A.1.3 Proof of Proposition 2.1 . . . 119
A.1.4 Proof of Proposition 2.2 . . . 120
A.1.5 Proof of Proposition 2.3 . . . 121
A.1.6 Proof of Proposition 2.4 . . . 122
A.1.7 Proof of Proposition 2.5 . . . 123
A.2 Proof of Results in Chapter 3 . . . 125
A.2.1 Proof of Proposition 3.1 . . . 129
A.2.2 Proof of Proposition 3.2 . . . 130
A.2.3 Proof of Proposition 3.3 . . . 131
A.2.4 Proof of Proposition 3.4 . . . 132
A.2.5 Proof of Proposition 3.5 . . . 133
A.2.6 Proof of Proposition 3.6 . . . 136
A.2.7 Proof of Proposition 3.7 . . . 137
A.2.8 Proof of Proposition 3.8 . . . 138
A.2.9 Proof of Proposition 3.9 . . . 139
A.2.10 Proof of Proposition 3.10 . . . 139
A.2.11 Proof of Proposition 3.11 . . . 141
A.3 Proof of Results in Chapter 4 . . . 146
A.3.1 Proof of Lemma 4.1 . . . 146
A.3.2 Proof of Lemma 4.2 . . . 147
A.3.3 Proof of Lemma 4.4 . . . 147
A.3.4 Proof of Lemma 4.5 . . . 148
A.3.5 Proof of Proposition 4.1 . . . 149
A.3.6 Proof of Proposition 4.2 . . . 150
A.3.7 Proof of Proposition 4.3 . . . 151

Bibliography 152

# List of Figures


2.1 Timeline of the Model . . . . . 19

2.2 Probability $C_F$ offered in $f$-Frame Equilibrium . . . . . 30

2.3 Equilibrium Probability of Exiting the Frame, $\rho(f)$ . . . . . 32

2.4 Optimal Frame as a Function of $\kappa$ . . . . . 36

2.5 Cognitive Strategy Given Optimal Frame, $f^*$ . . . . . 38

2.6 Total Surplus as a Function of $\kappa$ . . . . . 40

2.7 Comparing Cognitive Strategies with Screening and Without . . . . . 43

3.1 Timeline of the Model . . . . . 59

3.2 Complete Goal Equilibria . . . . . 66

3.3 Optimal Complete Goals . . . . . 69

3.4 Optimal Incomplete Goals . . . . . 73

3.5 Optimal Goals . . . . . 77

3.6 Deviation from goal in state $c = c_H$ as a function of $\beta$ . . . . . 80

3.7 Willingness-to-pay for full commitment as a function of $\beta$ with $q = 1/2$ 83

# Chapter 1

# Introduction

This thesis develops a set of theoretical models that investigate two behavioral phenomena: the fact that an individual's beliefs can be distorted from the truth, and that individuals use different modes of cognition in order to make decisions, where transition between these modes can occur as a function of the decision problem at hand.

There are many settings in which an individuals beliefs can be distorted away from the truth. This thesis will look at two such settings. The first is an interpersonal environment where a contract designer can distort the beliefs of a consumer through how he or she presents or frames information about the product or contract that is being provided. The second is an intrapersonal setting where an individual can distort her own beliefs through the use of different types of goals, and may find it advantageous to do so. In these sections, the trade-offs associated with belief distortion will be made explicit and a characterization of the optimal extent to which beliefs are distorted will be provided.

The extent to which an individual is cognitively active is an important determinant of whether the decision she makes is in her own self interest. When someone is in a state of low cognition, they may be impacted by forces that are attempting to distort their understanding of the problem at hand, or hold incorrect beliefs over how information is being generated. Thus, being in a state of low cognition may increase susceptibility to exploitation. In contrast, when an individual is in a state of high cognition, they may be able to question the source of their beliefs and identify the true beliefs they should hold. However, cognition is burdensome and so being in a state of high cognition comes at some cognitive cost. This thesis will use a model of costly cognition of this form in two settings. The first is in conjunction with the theory of belief distortion through framing effects, where costly cognition serves as a limitation on the extent to which a principal can distort the beliefs of an agent. The second is in a communication setting, where only through costly cognition can an individual correctly perceive the correlation structure through which information has been generated. In what follows, a more detailed description of the content of each chapter is provided.

Chapter 2 investigates a setting where both beliefs distortions and endogenous modes of cognition are important considerations. In particular, this chapter focuses on the interplay between cognition and the framing of contracts in a contract-design setting. The model presented provides an argument that the extent to which an individual is invested in cognition is an important moderator of framing effects. The details are as follows. A principal provides a consumption contract to an agent in an uncertain environment. The principal also chooses how to frame the contract which, if the agent is susceptible to the frame, distorts her beliefs from the truth. With cognitive effort, however, the agent is able to question the source of her beliefs, escape the frame, and correctly value the contract. The key contribution of this work is that

the likelihood that the agent is in either a low or high cognitive state is determined in equilibrium. This is achieved by balancing the marginal benefits of making a decision using the true contract valuation (and avoiding exploitation) against some cognitive costs. This generates a trade-off for the principal: distort beliefs further from the truth in order to earn larger profits on the low-cognition agent, at the cost of increased investment in cognition. By allowing the cognitive state of the agent to be endogenously determined, the model provides an intuitive set of comparative statics which depend crucially on equilibrium effects. Most interestingly, it is shown that, as the cost of cognition increases, the agent may actually invest more in cognition as a response to increased exploitation in the form of more distorted beliefs. The model is also extended to allow for cognition-based screening, in which it is shown that (a) the principal always exploits frame-susceptible agents more when able to screen, and (b) allowing for screening maximizes total surplus if and only if the marginal cost of cognition is sufficiently large.

Chapter 3 focuses on an intrapersonal setting in which an agent is able to distort her own beliefs through the use of different types of goals, by affecting the set of states she is aware of at the time of decision-making. The motivation for this work is to answer the question "why do individuals often, and repeatedly, set unrealistic goals for themselves?" The model provided jointly predicts that goals (a) do help to improve individual outcomes and (b) may be deviated from in equilibrium. It departs from the rational-expectations reference-point formation theory of Kőszegi and Rabin (2006) by introducing the notion of endogenous awareness: the set of states of the world that the individual considers possible at the *planning* stage may be different from those she is aware of at the *doing* stage, and the differences are determined by the type of goal utilized. More incomplete (complete) goals induce lower (greater) degrees of awareness. A two-state version of the model is first explored. It is shown

that incomplete goals (which will be deviated from in equilibrium) are optimal for individuals with relatively low levels of self-control. When such goals are optimal, the doing self of the agent is *over-optimistic* in the sense that she is induced to believe that only the most valuable state of nature will realize. Given this, goal deviation involves the individual systematically falling short of her self-set goals. Incomplete goals, however, are blunt self-regulatory devices and are shown to become dominated by more complete goals as the self-control problem dissipates. It is shown that the magnitude of goal deviation is non-monotonic in the self-control problem of the individual. In addition, the willingness-to-pay for external commitment is minimized when the magnitude of goal deviation is maximized, which suggests that one must be careful using the extent of goal deviation to determine the need for paternalistic intervention. Results are shown to be robust to (i) allowing for goal abstention, and (ii) allowing for more general, finite state-spaces.

Chapter 4 embeds the theory of endogenous cognitive-modes developed in Chapter 2 in a communication setting. The motivation for this is to write down a theoretical model which jointly captures that (a) individuals can be either naive or sophisticated with respect to their ability to understand the information, (b) that transmitters of information may attempt to take advantage of such naiveté, and (c) individuals should be able to endogenously transition between naiveté and sophistication as a function of the context in which information is generated. The framework utilized is a classic model of cheap-talk communication in a dynamic setting, with the addition that the receiver can either be sophisticated (i.e. perfectly understand the correlation between messages and states) or naive (have an incorrect perception of this joint distribution). Whether the receiver ends up naive or sophisticated is modeled as a function of costly investment in cognition which is determined as a best-response through the following trade-off: being in a state of high-cognition (sophisticated)

permits the receiver to make a more informed decision but this comes at a cognitive cost. Conditions are derived such that the sender will choose to truthfully reveal information in equilibrium with positive probability only when the receiver is cognitively bounded. This implies that bounded cognition can be a positive force for information revelation: the benchmark version of the model in which the receiver can observe the true joint distribution of signals and states does not permit any truthful revelation of information. In addition, a paradox of cognition is established whereby the sender may actually reveal information truthfully to a *greater* extent as the cost of cognition *increases*.

# Chapter 2

# Framing and Cognition in Contract Design

## 2.1 Introduction

The motivation for the work in this chapter is based on two fundamental ideas. First, the way in which a firm presents or *frames* information about the product they are offering can influence a consumer's valuation of said product. Second, the extent to which a firm is able to lead an individual to over-value a product or contract through the use of framing effects is limited. In this paper, we will argue that investment in cognition is an important moderator of the extent to which framing effects can be used for exploitation. This is investigated in the context of an uncertain contracting environment, where a principal offers a state-contingent consumption plan to an agent.

The class of framing effects we explore in this paper are those that result in the agent holding distorted beliefs. Formally, the agent should hold some beliefs $P$, but the framing of the contract induces the agent to hold some alternative set of beliefs,

$f$. This general treatment incorporates a number of theories of belief distortion in which an individual does not arrive at the correct posterior (in the Bayesian sense). It includes interpretations in which no true information has been transmitted ($P$ is the prior) but the agent mistakenly elicits information from the frame, as well as those in which information is transmitted, but the frame induces the use of a biased updating rule.

There are many examples of frames that induce belief distortions. First, advertising is a key strategy firms use in order to impact the beliefs of consumers. For example, the framing of a product (through the lens of an advertisement) may induce stereotypical thinking (Bordalo et al. (2016)), or result in salience effects and updating through use of the representativeness heuristic (Gennaioli and Shleifer (2010), Bordalo, Gennaioli, and Shleifer (2012), Bordalo, Gennaioli, and Shleifer (2013)), or lead an individual to form categories and use spurious analogies (Mullainathan, Schwartzstein, and Shleifer (2008)). Second, the principal may make use of examples or use different font-sizes (contract terms expressed in the 'fine-print' versus those that are not) to induce the agent to overweight particular states. Here, we will assume that the principal has access to a wide range of framing technologies that can induce a set of distorted beliefs. The reason for this is to show that, even with an unlimited capacity to exploit the agent, the interactions with the agent's cognition-state may limit the incentives to do so.

The process through which the frame is transformed into choice is determined by the cognitive-state of the agent. Specifically, we assume that the agent is either in a state of low cognition or a state of high cognition. In the low-cognitive state, the agent is unable to observe the impact of the frame and she makes decisions using the distorted beliefs. In contrast, if the agent is in a state of high-cognition, she is able

to question the motivations of the principal, recognize the impact of framing, and evaluate the contract using the true beliefs.

One can interpret the cognitive-state of the agent as the extent to which she is thinking about or paying attention to the subtle, but relevant factors that affect her decision-making. When the agent is in a low-cognitive state, her decision-making is automatic and no attention is paid to whether or not she has been influenced by framing. In contrast, when the agent is in the high-cognitive state, then her decision-making process is more controlled: she is able to think about the problem she is facing, pay attention to the way the contract is framed, and recognize the impact it has on her beliefs. In this sense, the model of cognition relates to the dual-process theories in the psychology literature; most famously the theory of System 1 (automatic decision-making) and System 2 (controlled decision-making) discussed in Kahneman (2011).

A key contribution of the paper is that we allow for the probability that the agent is in either cognitive state to be endogenously determined in equilibrium: the marginal gain in utility of moving from a low-cognitive state to a high-cognitive state is balanced against the marginal costs of cognition. This allows for different framing effects to trigger varying degrees of suspicion in the agent. When the agent becomes suspicious, a high cognitive-state is triggered and she is able to escape the effects of framing. By being in a state of constant suspicion, however, is cognitively taxing and, as such, the agent will sometimes allow herself to be susceptible to the frame. In equilibrium, we will be able to provide theoretical predictions of the factors that trigger more automatic or controlled decision-making processes within the individual.

This bound on the agent's rationality permits exploitation by the principal. If able to identify a framed type, the principal will *over-provide* consumption in states that are overweighted under the frame and will *under-provide* consumption in states that are underweighted under the frame. This allows him to charge a higher price for provision to the agent in the low-cognitive state, while providing the same amount of 'true' expected consumption that is offered in the absence of framing effects. Such distortions are shown to be profitable to the principal.

We first proceed by investigating a baseline version of the model in which the principal is restricted to offer only a single contract. The objective is to display the importance of allowing for the cognitive-state of the agent to be endogenously determined in equilibrium. We show that the cognitive strategy of the agent is increasing in a measure of the extent to which the frame distorts beliefs away from the truth. At the same time, the profit that the principal is able to make on the frame-susceptible agent-type is increasing in this same measure. This implies that the principal faces the following equilibrium trade-off: distort beliefs more in order to increase profits on agents in a low-cognitive state at the cost of decreasing the relative proportion of such agents in the market.

The optimal balancing of this trade-off is explicitly solved for. It is shown that the amount of exploitation involved in the optimal frame is increasing in the agent's marginal cost of cognition. This exploitation, however, is often far from maximal due to limitations imposed by allowing the cognition strategy of the agent to vary with the frame. Indeed, when the marginal cost of cognition is sufficiently small, the principal may abstain from using framing effects in equilibrium.

The model provides an intuitive set of comparative statics, which depend crucially on the model's equilibrium effects. In particular, it is shown that the individual may actually invest more in cognition as the marginal cost of cognition increases. This is due to the fact that the individual balances the increasing cognitive costs (which depress cognition) against the equilibrium effect of increased exploited (which incentivizes cognitive investment). When both forces are present then, under the assumed functional forms, the latter effect dominates and investment in cognition increases in its own cost. Said in another way, with an increase in cognitive costs, the principal optimizes on frame choice by deciding to sell less contracts, which is more than compensated for by an increase in per-contract sale profits.

The model is extended to allow the principal to engage in cognition-based screening by offering a menu of contracts. This generates a similar set of comparative statics from the baseline case, where only a single contract was offered. There are, however, some important differences. By comparing across the two settings it is shown that (a) the frame the principal chooses is *always* more exploitative when he is able to screen; and (b) total surplus is greater when screening is allowed if and only if the marginal cost of cognition is sufficiently large.

The first finding is intuitive: by being able to screen, the principal is now able to extract surplus from agents in a high state of cognition, who were not participating in the market in the baseline version of the model. This implies that having the agent transition into a high-cognitive state is now less costly and, as such, the principal is willing to choose a more exploitative frame.

The second finding, in contrast, requires a more nuanced explanation. In the 'no-screen' treatment, when the marginal cost of cognition is low, it is an equilibrium

for the principal to not employ a framing technology, which implies that the efficient market outcome is achievable. In contrast, the principal *always* distorts beliefs when screening is allowed. As the cost of cognition increases, however, the principal will eventually find it optimal to maximally exploit the agent, irrespective of whether screening is allowed or not. Then, the agent in each cognitive-state is equally well off across the two market settings. Since the principal can also earn profits on agents that hold true beliefs in the screening condition, it follows that total surplus is higher in this setting. This has important implications for regulators, depending on whether they are focused solely on minimizing consumer exploitation (such as consumer watchdogs) or have an objective to maximize market efficiency (measured by total surplus).

## 2.2 Related Literature

**Framing:** There is large literature on framing effects and their impacts in settings of economic interest. Salant and Rubinstein (2008) provide a decision-theoretic framework where the domain of choice is expanded to include the way that choice-objects are framed. Formally, the authors model choice as a function $c(A, f)$ where $A$ is the set of choice-alternatives, $c(A, f) \in A$ is the choice from $A$, and $f$ is the frame under which the decision is made.

In their paper, the authors specify a number of interesting forms that $f$ may take. In this paper, we focus on a particular class of frames (those that induce belief distortions) but simultaneously provide a general treatment of frames within this class. We also provide a novel theory of how such frames are transformed into choice as a function of the extent to which an individual has invested in cognition.

Benkert and Netzer (2016) investigate a setting in which an individual is prone to mistakes when making choices. The authors identify conditions on classes of behavioral preferences for which the framing of decision problems can positively impact the agent. Ahn and Ergin (2010) also provide a choice-theoretic model in which the subjective likelihood of a contingency depends on how it is described, which is a special case of the set of belief-distortions allowed here. In contrast to these papers, we analyze a strategic setting in which a principal selects the frame, rather than focusing on precisely how framing affects choice behavior. We also provide a mechanism through which the individual can escape the effects of frames through costly cognition.

The work here is most closely related to the work of Salant and Siegel (2016). In their model, a frame directly affects the preferences of the individual and the limitation the principal faces is the fact that the frame will wear off and the individual will be able to return the product if she was exploited. Instead, in this paper the frame specifically impacts the *beliefs* of the agent, while leaving her state-by-state preferences unchanged. Moreover, the mediator of framing effects is cognition, which affects the ability for the principal to exploit the agent *ex-ante* (by seeing through the frame and not purchasing the product) rather than *ex-post* (through the ability to return the product).

Both Piccione and Spiegler (2012) and Spiegler (2014) investigate a complete-information competitive setting where frames, which are the result of 'marketing messages' chosen by the firms, decrease the consumer's ability to make comparisons. In contrast, this paper investigates a monopolistic environment in which there is asymmetric information, as the principal does not observe the agent's cognitive state. It also analyzes an entirely distinct notion of framing from the one in their paper.

Finally, Caplin and Martin (2012) investigate a model in which the individual can observe some properties of the problem (for example, where an object appears in a list) but not the payoffs to choosing an option. The individual uses a prior to correlate the position in the list and the payoff to form posterior beliefs regarding value. Similar to this paper, framing effects arise endogenously in their model, but for very distinct reasons. In their paper, this arises due to the fact that the individual must pay a cost to attain information regarding the states of the world (i.e. the payoff). In contrast, here the agent is able to correctly value the contract in each state of the world, but is limited due to the fact that it is costly to question whether her beliefs have been perturbed. Moreover, again we are focused on a strategic setting where the principal chooses the frame.

**Awareness and Cognition in Contracts:** The paper is also related to the literature on awareness and its implications for contract design. In Filiz-Ozbay (2012), the agent is aware of only a subset of the states of the world that the principal is aware of, and the principal must decide whether to increase the awareness of the agent (i.e. move her beliefs closer to the truth). In contrast, in this paper the principal and the agent start out with identical beliefs and the principal must decide whether to distort the agent's beliefs further from the truth. For a specific class of belief distortions based on awareness, this means that the principal must decide whether or not to *reduce*, rather than increase, the awareness set of the agent. In addition, the agent in this paper is given the cognitive ability to become 'aware', without the principal's aid.

In Von Thadden and Zhao (2012), the agent is unaware of all the actions available at her disposal and, as long as the principal chooses an incomplete contract (modeled as a non-state-contingent wage), this unawareness is maintained and a default option is chosen. It is shown that maintaining unawareness can be optimal

for the principal as awareness of the action space introduces incentive constraints that must be satisfied and which transfer rents from the principal to the agent. In contrast, the current work is focused on the case in which the beliefs of the agent are distorted, rather than limiting his awareness of the action space. Again, the agent also has the ability to become 'aware' of a distortion through her cognition, whereas in Von Thadden and Zhao (2012) this can only incur with principal intervention.

In the literature on cognition and contracts, this work is most closely related to that of Tirole (2009). In his paper, a buyer and a seller bargain over the surplus generated by a contract, which may or may not be the surplus-maximizing contract. With an investment in cognition, the players may discover that there is a better contract, because it is more complete, with more aptly designed covenants. Similarly, in the current paper the level of cognition is endogenously determined in equilibrium. The modeling approaches are distinct, however, in a number of dimensions. First, in Tirole (2009), the default contract is exogenously given. In contrast, in this paper this contract is chosen by the principal, and its 'default' value is a function of the framing strategy. Second, if a more suitable contract exists, then it is indescribable in his setting, unless discovered with cognitive effort. Here, any contract can be described ex-ante: the cognitive discovery of the agent is related to whether her beliefs have been distorted or not.

**Behavioral Contract Theory:** The work is also related to the more general literature on behavioral contract theory (see Kőszegi (2014) for an in-depth literature review). There are a number of papers that investigate models were a fraction of the population is naive in some sense. Some different forms of this include settings where some consumers are naive and misperceive the price of a good to be lower than its true price (Gabaix and Laibson (2006), Armstrong and Vickers (2012)), where indi-

viduals are naive regarding their self-control problems (DellaVigna and Malmendier (2004), Heidhues and Koszegi (2010)), or where individuals naively interpret signals in communication games (Chen (2011), Ottaviani and Squintani (2006)). In all of these papers, whether the agent is naive or sophisticated is fixed aspect or a 'type'. Instead, it may be more reasonable to model individuals as being able to transition between naiveté and sophistication, and to have this depend on endogenous cognitive states.

## 2.3 Model

We first present the formal details of the model, then discuss the model's key assumptions in further detail.

### 2.3.1 Primitives

There is a finite state-space, $\Omega = \{\omega_1, \ldots, \omega_n\}$, with prior $P$ governing uncertainty over the states. In an abuse of notation, we write $P(\omega)$ for $P(\{\omega\})$. There is a principal (he) and an agent (she).

**Principal's Strategy:** The principal designs a contract for the agent with the objective of maximizing expected profits. Denote by $C = (q, p)$ an arbitrary contract, which is the combination of a consumption function, $q : \Omega \to [0, \infty)$ and an unconditional price for the contract, $p \geq 0$. The principal can produce $x$ units of consumption at cost $C(x) = cx^2/2$ where $c > 0$ is constant across $\Omega$.

In addition to designing the contract, the principal also selects a framing of said contract. We model a frame as a function $f : \Delta(\Omega) \to \Delta(\Omega)$ that maps the prior, $P$, to an alternative set of beliefs, $f(P)$. We can identify a frame, $f$, with the posterior

belief it induces $f(P) \in \Delta(\Omega)$, where we write $f_P \equiv f(P)$. With the interpretation that $P$ is the agent's prior, the contract's framing *distorts the beliefs* of the agent, even though no actual information has been transmitted. Let $\mathcal{F} \subset \Delta(\Omega)$ denote the set of feasible posteriors that can be induced by some frame $f$. For the majority of the analysis, we focus on the case in which $\mathcal{F} = \Delta(\Omega)$ so that the principal can induce the agent to hold *any* beliefs over the state-space with a suitably chosen frame.

**Agent Preferences:** The individual has a vNM utility index $u(x) = \theta x - p$, where $x$ is a level of consumption received and $p$ is the price paid for that level of consumption. The individual has an outside option $q_0 : \Omega \to [0, \infty)$, which, given the assumption that the consumption utility is linear, can be normalized to $q_0(\omega) = 0$ for all $\omega \in \Omega$.

**Cognitive States:** The individual is in one of two cognitive-states: a low-cognition state (denoted by $L$) or a high-cognitive state (denoted by $H$). We denote by $\rho \in [0, 1]$ the probability that the agent will be in cognitive-state $H$. Importantly, $\rho$ will be endogenously determined in equilibrium (to be discussed subsequently).

We assume that the agent is able to understand the terms of the contract, irrespective of her cognitive-state. Formally, for a contract, $C = (q, p)$, she is always able to correctly compute the utility value of the contract in each state of the world: $u(\omega) = \theta q(\omega) - p$. Her cognitive-state, however, will be an important determinant of her ability to recognize the impact that the frame has on her beliefs.

Specifically, suppose that the agent is in cognitive-state $L$. In this state, the agent is unable to observe the impact that the framing of the contract, $f$ has on her beliefs. Given this, she will use the distorted beliefs induced by the frame, $f_P$, as weights

in the computation of her expected utility. Thus, an individual in cognitive-state $L$ values a contract $C = (q, p)$ by

$$U_L(C, f) = \sum_{\omega \in \Omega} \theta q(\omega) f_P(\omega) - p. \tag{2.1}$$

She *overvalues* consumption in states of the world with $f_P(\omega) > P(\omega)$ and *undervalues* consumption in states with $f_P(\omega) < P(\omega)$. This belief distortion will ensure that the principal is able to exploit frame-susceptible agents for profitable gain.

In contrast, if the individual is in a high-cognitive state, she is able to question the motivations of the principal, recognize that her beliefs are distorted by the contract's framing, and subsequently 'escape the frame'. Hence, she will use the prior $P$ to evaluate a contract $C = (q, p)$:

$$U_H(C, f) = \sum_{\omega \in \Omega} \theta q(\omega) P(\omega) - p \tag{2.2}$$

which is independent of the frame $f$. Given this, the principal is unable to induce an agent in a state of high-cognition to buy a contract that is dominated by her outside option.

**Endogenous Cognition:** A key assumption in this paper is that the probability that the agent is in a particular cognitive-state ($H$ or $L$) will be endogenously determined in equilibrium. Let $\rho$ denote the probability that the agent is in cognitive-state $H$. We will sometimes call $\rho$ the agent's cognitive strategy. Then, the equilibrium cognitive strategy of the agent, $\rho^*$, is determined as follows. Fix a frame, $f_P \in \mathcal{F}$ and contract $C = (q, p)$. Let $a_i \in \{0, 1\}$ denote the action of the agent in cognitive-state $i \in \{L, H\}$, where $a = 1$ corresponds to purchasing contract $C$ and $a = 0$ implies that agent does not participate. Then, $\rho^*$ is given by the solution to the following

optimization problem:

$$\max_{\rho} \Big\{ \rho U_R(a_H|C,f) + (1-\rho) U_R(a_L|C,f) - T(\rho) \Big\} \tag{2.3}$$

where $T(\cdot)$ is interpreted as the cognitive costs (in terms of utility) of increasing the probability of triggering high cognition. We impose the functional form $T(\rho) = \kappa\rho^2/2$, where $\kappa > 0$ can be interpreted to be the marginal cost of cognition. Notice that (a) $T(0) = 0$ (no investment in cognition is free), (b) $T(\cdot)$ is strictly increasing in $\rho$ (the higher the likelihood of rationality, the higher the cognitive cost), and (c) $T$ is strictly convex (marginal investments in cognition are increasingly costly).

An implicit assumption made in equation (2.3) is that $\rho$ is determined using the *true* preferences of the individual, $U_R$. Thus, $\rho$ can be interpreted as the ex-ante equilibrium level of exploitation that the agent is willing to endure, balanced against how costly high-level cognition is. A suitable interpretation of agent's cognitive strategy is that it is an *automatic process* that triggers either high or low cognition, as a function of the agent's environment. By requiring that $\rho$ is determined by this optimization problem, we can impose some natural structure on how cognitive-states are triggered by the situation at hand.

**Equilibrium:** The model introduces some nuanced issues related to the timing. More specifically, one must make an assumption regarding what the agent has observed at the point in which the cognitive-state is triggered. In the baseline version of the model, we utilize the following timing (timeline provided in Figure 1). The principal selects his strategy, which is the combination of a contract, $C$, and a framing of that contract, $f_P$. The agent first observes $f$, at which point her cognitive-state is selected (according to $\rho$). The agent then observes $C$ and decides

whether or not to purchase $C$ as a function of her realized beliefs.

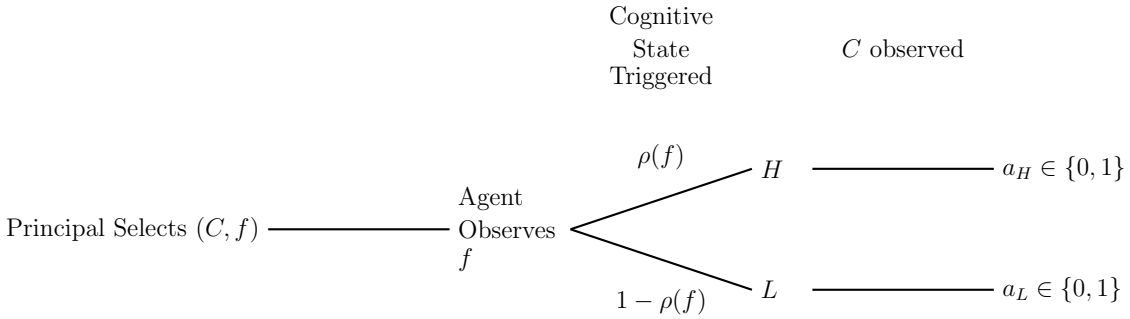


Figure 2.1: Timeline of the Model

With this timing, the agent has not *directly* observed the contract $C$ and her subconscious must instead make a prediction regarding what that contract will be (some $\hat{C}$) in order to determine her cognitive strategy. In equilibrium, her prediction will be correct: $\hat{C} = C$. In this sense, given a frame $f_P$, one can think of $C$ and $\rho$ as being determined *simultaneously* (as a Nash equilibrium). Then, the principal will only have *direct* control over the agent's cognitive-type through the choice of framing device. We will discuss the differences between this formulation and alternative timing assumptions in the following section.

Given the timing of the model, in order to search for the equilibrium we can proceed in two steps: first, fix a frame $f_P$ and solve for the equilibrium contract $C^*(f_P)$ and cognitive strategy, $\rho(f_P)$; second choose the framing strategy that maximizes expected profits across all frames. The first step, labeled an $f$-frame equilibrium, is defined below.

**Definition 2.1.** *An $f$-frame equilibrium consists of a transition probability, $\rho(f)$, a contract $C^*(f) = (q_f^*, p^*(f))$ and actions $a_i^*(f) \in \{0,1\}$ for $i = L, H$, such that*

*(1)* $a_i^*(f) = 1$ *if and only if* $U_i(C^*(f), f) \geq 0$.

*(2)* $C^*(f)$ *solves*

$$\max_{C=(q,p)} \left\{ (\rho(f)a_H^*(f) + (1 - \rho(f)a_L^*(f)) \left[ p - c \sum_{\omega \in \Omega} P(\omega) \frac{q(\omega)^2}{2} \right] \right\}$$

*(3)* $\rho(f)$ *solves*

$$\max_{\rho} \Big\{ \rho U_R(a_H^*(f)|C^*(f), f) + (1 - \rho) U_R(a_L^*(f)|C^*(f), f) - T(\rho) \Big\}.$$

The first condition is that the individual makes a decision that is (subjectively) individually-rational, given her cognitive-state. The second condition is that the principal selects the contract to maximize expected profits given the agents' cognitive strategy and resulting purchasing decisions. The final condition is the equilibrium determination of $\rho(f)$, as previously discussed.

Given each $f$-frame equilibrium, the principal then chooses the optimal frame: $f^*$ solves:

$$\max_{f_P \in \mathcal{F}} \left\{ (\rho(f)a_H^*(f) + (1 - \rho(f))a_L^*(f)) \left[ p^*(f) - c \sum \omega P(\omega) \frac{q_f^*(\omega)^2}{2} \right] \right\}. \tag{2.4}$$

### 2.3.2 Discussion of Model's Assumptions

The benchmark model described above embodies four key assumptions, which we discuss in turn: the particular types of framing effects allowed for in the model, the concept of endogenous cognition, the assumption that the principal can only use a single contract, and the timing of events in the model.

**Framing:** The notion of a *framing effect* can be distilled into the following fundamental distortion: the individual is induced to use beliefs $f_P$ to make decisions when the true beliefs are given by $P$. The interpretation offered in the previous section is that $P$ represents the prior and the principal utilizes an uninformative signal which the agent mistakes for information that induces posterior $f$. Some examples of framing effects of this form include categorization effects which leads to the use of spurious analogies (Mullainathan, Schwartzstein, and Shleifer (2008)) or endogenous awareness, where the individual focuses, and conditions, on a highlighted subset of the states (Young (2017)).

More generally, there is nothing that requires $P$ to be the prior. Instead, given some actual information, $P$ could represent the *true* Bayesian update the individual should use while $f_P$ is the result of some biased updating procedure. For example, the principal may be choosing a frame that induces stereotypical thinking (Bordalo et al. (2016)) or induces the agent to use a form of the representativeness heuristic (Gennaioli and Shleifer (2010), Bordalo, Gennaioli, and Shleifer (2012), Bordalo, Gennaioli, and Shleifer (2013)). An investment in cognition in these cases would involve the individual recognizing that her updating procedure is biased, and using the true information received to update (in the Bayesian sense).

By calling a frame the resulting beliefs that the framed agent uses to perform valuations, we remain agnostic on what the specific procedure inducing the belief distortion is. Thus, this is a relatively general model of belief-based framing that incorporates many of these models that focus more on *how* belief distortions can occur. The reason for taking such a general approach is to show that cognitive responses have strong moderating effects on the principal's ability to frame, even

when he has access to a wide range of belief-distorting technologies.

It should be noted that by focusing on framing effects that result from belief distortions, we do exclude, *a priori*, alternative framing effects that operate directly on the individual's preferences. These include those that affect how the individual values losses versus gains (Kahneman and Tversky (1981)) or those that impact the *feelings* about risk (by perturbing the coefficient of risk aversion, for example). A more general model would have framing directly impact the expected value of the contract (whether through beliefs or preferences). We, however, focus on the case in which framing can only distort beliefs because (a) there is large amount of empirical evidence that such belief-distortions persist, (b) it incorporates a number of theories of biased beliefs, (c) it allows us to isolate the impact such effects have on state-by-state contract design, and (d) to ensure there is a natural limitation on the extent to which the principal can exploit the agent.

**Endogenous Cognition:** The agent's cognitive strategy is modeled as a *probability* with which she will be susceptible to framing effects. This approach is taken to fit with realistic view that an individual is not always in only a single cognitive-state, but rather is likely to transition between these cognitive states, depending on the problem she faces. In addition, by allowing for this to be determined endogenously, we are also able to provide predictions for regarding which situations one should expect to observe an individual transitioning into a higher cognitive state.

From a modeling standpoint, the fact that the individual can be in different mental-states is a key assumption that generates the trade-offs in the model. Specifically, it ensures that there is asymmetric information: the principal will not be able to identify whether the agent has adopted distorted beliefs or escaped the frame and

will have to design the optimal contract accordingly. The fact that this is determined endogenously, *after* observing the framing of the contract, is also important for generating the model's key findings. In particular, if the cognitive-state of the agent was determined by some exogenous procedure, then the principal would always wish to use a maximally exploitative frame.

**Single Contract:** In the baseline version of the model, we focus on the case in which the principal is restricted to offer a single contract. This is for a number of reasons. First, being faced with a menu of contracts may exogenously interact with the agent's cognitive state. For example, if the principal offers two contracts from which the agent has to choose, then the fact that the individual is forced to compare these two contracts may, in and of itself, trigger a controlled decision-making process. As a by-product, the agent may also identify that her beliefs are distorted due to the framing of the contract and adjust accordingly. If this were the case, it would be strictly optimal for the principal to use a single contract. The extent to which the number of choice objects influences cognitive investment is an empirical question that requires further investigation.

Second, following the analysis of the single-contract case, we do extend the model to allow for the principal to present a menu of contracts and show that this offers similar predictions regarding optimal frame choice, with some nuanced differences between equilibrium cognitive investment and market efficiency. This has important implications for whether a regulator (with a particular set of objectives) should seek to identify whether the extent of cognition-based screening in a market.

**Timing:** As discussed in the previous section, we assume that the agent has only observed the framing strategy of the principal at the time her cognitive-state is

triggered. Instead, the cognitive-strategy and the optimal contract are determined simultaneously in equilibrium. This seems to be a reasonable assumption in a number of settings. For example, if the frame is based on font size or the amount of information contained in the 'fine-print' of the contract, then it seems reasonable that the individual will observe and process this before identifying the contract's value determined by its terms. Similarly, if the frame is in the form of an advertisement, the agent may observe this before identifying the product's value[1].

It is the case, however, that there are alternative formulations of model's timing that one may argue are suitable. For example, the cognitive-state of the agent may be triggered after observing both the frame and the contract. This would imply that the principal can directly impact the agent's cognitive strategy by varying the contract.

The particular timing chosen is utilized as it leads to a clear characterization of the equilibrium and a sharp and intuitive set of comparative statics. The implications of alternative formulations, however, should be explored.

## 2.4 Single-Contract Offered

In this section, we characterize the equilibrium variables of interest for the baseline model described in the previous section. We first investigate the version of the model in which the principal is only able to offer a single contract, after which we compare the results to the setting in which the principal is able to screen on the agent's cognitive type (solved in Section 5).

[1]Indeed, advertisements are often abstract and offer no specific information regarding the product that is being sold. Instead, one may think that the goal of such ads are to distort the agent's beliefs regarding the state (for example, the quality of the firm itself).

In each case, we first proceed by characterizing the $f$-frame equilibrium for each frame $f$. It is shown that the key variables of interest (the equilibrium cognitive strategy of the agent $\rho(f)$ and the per-unit profit on each individual contract sale) are functions of a useful measure of divergence between the posterior induced by the frame and the prior distribution, which allows the principal's frame-optimization problem to be simplified greatly. Given this, the optimal frame is derived and comparative statics are provided for measures of consumer exploitation and welfare.

### 2.4.1 $f$-Frame Equilibria

For this section, we fix a frame $f$ which induces distorted belief $f_P$ if the agent is susceptible to the frame. From the perspective of the principal, the cognitive-state of the agent is fixed. Therefore, we will sometimes refer to an individual that is in a high-cognitive state as the *rational type* and an individual that is in a low-cognitive state as the *framed type*.

We begin with the following lemma, which shows that the only contracts ever offered in equilibrium are those which are optimal if the cognitive-state of the agent were observable.

**Lemma 2.1.** *Define the* ***rational contract****,* $C_R$*, to be*

$$q_R(\omega) = \frac{\theta}{c}, \quad p_R = \frac{\theta^2}{c}$$

*and the* ***framed-optimal contract****,* $C_F$*, to be*

$$q_F(\omega) = \frac{\theta}{c}\frac{f_P(\omega)}{P(\omega)}, \quad p_F = \frac{\theta^2}{c}\sum_{\omega}\frac{f_P(\omega)^2}{P(\omega)}.$$

*Then, $C_R$ and $C_F$ are the only contracts ever offered by the principal in equilibrium. If $C_R$ is offered, then both agent-types purchase the contract and if $C_F$ is offered, then only type $H$ purchases the contract.*

*Proof.* See Appendix. ■

This lemma shows that the principal will only ever utilize one of two contracts in equilibrium: those which would be optimal if the principal could observe the cognitive-state of the agent. Comparing $C_F$ and $C_R$, we see that there is *over-provision* of consumption in $C_F$ relative to $C_R$ for states that have higher likelihood under the frame ($(f(\omega) > P(\omega))$ and there is *under-provision* of consumption in the framed contract for states with $f(\omega) < P(\omega)$. This allows the principal to charge a higher price (by a factor of $\sum_{\omega} f(\omega)^2/P(\omega) > 1$) for a consumption plan that offers the same level of expected consumption utility (i.e. $\sum_{\omega} P(\omega)q_F(\omega) = \sum_{\omega} P(\omega)q_R(\omega) = \theta/c$).

We now define a useful measure of the divergence between two probability distributions, which will capture the extent to which the agent's beliefs are distorted by frame, $f$.

**Definition 2.2.** *Let $X$ and $Y$ be two probability measures over $\Omega$. The $\chi^2$-divergence measure of $X$ from $Y$ is given by*

$$D(X||Y) = \sum_{\omega \in \Omega} Y(\omega) \left( \left( \frac{X(\omega)}{Y(\omega)} \right)^2 - 1 \right).$$

Let $f^1$ and $f^2$ denote two distinct frame-induced beliefs. Then, if $D(f^1_P||P) > D(f^2_P||P)$, we will say that frame $f^1$ is *more distorting* than $f^2$. It is useful to think of this divergence as a measure of the extent to which the individual is *surprised* by having the distorted beliefs resulting from the frame. It is will be shown that,

in equilibrium, the more surprised the agent is (the larger is $D(f_P||P)$), the more suspicious she is of her beliefs and, thus, the greater her incentive to invest in cognition.

The $\chi^2$-divergence measure falls into a class of divergence measures which can be more generally labeled $g$-divergence measures[2]. A $g$-divergence measure is one in which, there exists a convex function $g$ where $g(1) = 0$ such that

$$D_g(X||Y) = \sum_{\omega \in \Omega} Y(\omega) g\left(\frac{X(\omega)}{Y(\omega)}\right).$$

Examples of such measures of divergence include relative entropy (or the Kullback-Leibler divergence measure) in which $g(t) = t \log t$. The reason we use the $\chi^2$-divergence measure is that it will be an important quantity for describing the form of the equilibrium variables in an $f$-frame equilibrium, due to the functional forms assumptions.

The following example provides a concrete calculation of the $\chi^2$-divergence measure for a specific class of belief-distorting frames.

**Example 2.1.** *Suppose that the set of feasible frames is such that $f_P \in \mathcal{F}$ if and only if there exists a set $E \subset \Omega$ such that*

$$f_P(\omega) = \frac{P(\omega)}{\sum_{\omega \in \xi} P(\omega)}.$$

[2]Technically, the label given to this class of divergence measures is $f$-divergence measures. Given that this conflicts with the use of $f$ as a frame in this paper, the notation has been changed to avoid confusion.

*One can interpret this as a set of frames that result from the use of examples. The principal selects a set of states to highlight (or make an example of), $E \subset \Omega$. The framed-type considers only the example states when evaluating the contract. Alternatively, the agent in a low-cognitive state mistakenly believes the set of examples contain true information, causing her to condition on $E$ and update her beliefs. Such a set of frames is equivalent to notion of "endogenous awareness" introduced in Young (2017).*

*In this case, each frame $f_P$ can be associated with the set $E_f \subset \Omega$ it is generated from. The $\chi^2$-divergence in this case is given by*

$$D(f_P||P) = \sum_{\omega} P(\omega) \left[ \left( \frac{f_P(\omega)}{P(\omega)} \right)^2 - 1 \right] = \sum_{\omega \in E} \frac{P(\omega)}{P(E)^2} - 1 = \frac{1}{P(E)} - 1.$$

*Intuitively, the more unrealistic that set of examples are for representing the set of things that might occur to the agent, the more her beliefs are distorted, according to this measure.*

We now proceed to characterize the equilibrium for the a given frame, $f$. Proposition 2.1 shows that the optimal contract to offer a framed type does not satisfy the individual-rationality constraint of the rational type. Hence, the principal faces a trade-off in terms of which contract to offer: offer $C_R$ and have all types purchase the contract or offer $C_F$, make higher profits on each sale individual contract sale, but decrease the fraction of types purchasing the contract (to $1 - \rho(f)$). It is optimal to offer both contracts if and only if

$$\frac{\theta^2}{2c} > (1 - \rho(f)) \frac{\theta^2}{2c} \sum_{\omega} \frac{f_P(\omega)^2}{P(\omega)} \Leftrightarrow \rho(f) \geq \frac{D(f_P||P)}{1 + D(f_P||P)} \equiv \bar{\rho}(f). \tag{2.5}$$

Only when the proportion of rational types is sufficiently large is it optimal for the firm to offer the contract $C_R$. When the proportion of framed types is sufficiently large, the principal will take advantage of the agents in a low-cognitive state and offer the exploitative contract, $C_F$.

An important fact to observe is that $\rho(f)$ can not be strictly greater than $\bar{\rho}(f)$ in an $f$-frame equilibrium. To see this, note that if (2.5) holds strictly, then the principal offers the unframed-optimal contract, $C_R$. By Lemma 2.1, both cognitive types purchase this contract which implies that $U_H(a_H^*(f)|C_R, f) = U_H(a_L^*(f)|C_R, f)$ and the optimal cognitive strategy of the agent would be to have $\rho(f) = 0 < \bar{\rho}(f))$ (since cognition is costly). Hence, $\rho(f) > \bar{\rho}(f))$ can not be a part of any $f$-frame equilibrium.

It follows that the principal will either (a) only offer the framed-optimal contract $C_F$ (if $\rho(f) < \bar{\rho}(f)$) or (b) will be willing to mix between $C_F$ and $C_R$ (when $\rho(f) = \bar{\rho}(f)$). The following lemma provides the equilibrium probability that the principal offers the framed contract as a function of the agent's marginal cost of cognition, $\kappa$.

**Lemma 2.2.** *Fix frame $f$ and let $\bar{\kappa}(f) = \frac{\theta^2}{c}(1 + D(f_P||P))$. Let $\eta^*$ denote the equilibrium probability that the principal chooses contract $C_F$ and $1 - \eta^*$ denote the probability the principal chooses contract $C_R$. Then,*

$$\eta^* = \begin{cases} \frac{\kappa}{\bar{\kappa}(f)} & \text{if } \kappa \leq \bar{\kappa}(f) \\ 1 & \text{if } \kappa > \bar{\kappa}(f) \end{cases}$$

*Proof.* See Appendix. ■

Figure 1 provides a plot of $\eta^*(f)$ as a function of $\kappa$. As can be seen, the likelihood that the principal offers the framed-optimal contract, $C_F$, is strictly increasing in the agent's marginal cost of cognition. This is intuitive: an increase in $\kappa$ dis-incentivizes investment in cognition. In order to keep the agent's cognitive strategy constant (equal to $\bar{\rho}(f)$), the principal needs to increase the risk that she will be exploited. Thus, $\eta^*$ increases in $\kappa$.

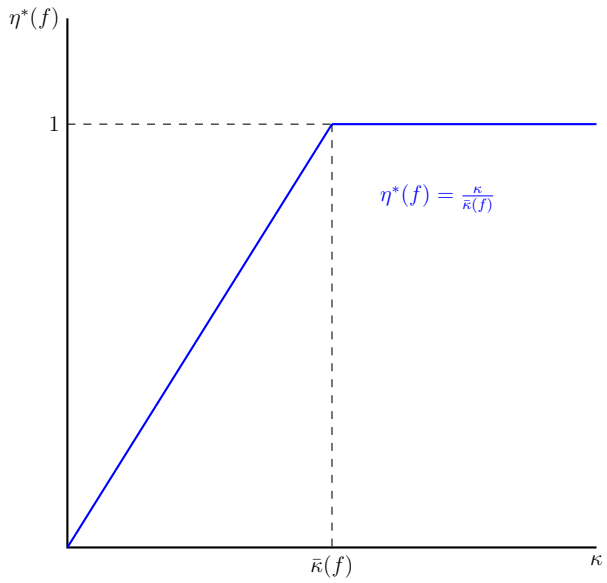


Figure 2.2: Probability $C_F$ offered in $f$-Frame Equilibrium

For $\kappa$ sufficiently small, the principal mixes between offering the framed-optimal contract and the rational contract. This is because when the marginal cost of cognition is low, the agent has large incentives to think. As previously argued, however, the probability the agent is in a high-cognitive state, $\rho(f)$, can not become so large that the principal wants to offer only the rational contract. Hence, in order to ensure $\rho(f)$ does not become "too large", the principal mixes between offering $C_F$ and $C_R$, where the probability that $C_R$ is offered is increasing as $\kappa$ decreases

(converges to one as the individual becomes "rational", at $\kappa = 0$).

On the other hand, when $\kappa$ is sufficiently large, the principal offers the framed-optimal contract with probability one. This is because $\rho(f)$ is sufficiently small for large $\kappa$ so as to ensure that $C_F$ strictly dominates offering $C_R$. The following proposition formalizes these ideas by providing a characterization of the equilibrium cognitive strategy for an arbitrary $f$.

**Proposition 2.1.** *Fix frame a frame $f$. Then,*

$$\rho(f) = \begin{cases} \bar{\rho}(f) & \text{if } \kappa \leq \bar{\kappa}(f) \\ \frac{\theta^2}{\kappa c} D(f_P||P) & \text{if } \kappa > \bar{\kappa}(f). \end{cases}$$

*$\rho(f) = 0$ only when $f_P = P$ and $\rho(f)$ is strictly increasing in $D(f_P||P)$.*

*Proof.* See Appendix. ■

Figure 2 provides a plot of $\rho(f)$ as a function of $\kappa$ for two different choices of frame, $f^1$ and $f^2$, such that $D(f_P^1||P) > D(f_P^2||P)$. As can be seen from Figure 2, for a given frame, $f$, the equilibrium probability that the individual engages in cognition is (weakly) decreasing in the marginal cost of cognition, which is intuitive. Moreover, it shows that $D(f_P||P)$ is not only a suitable measure of belief distortion, but also constitutes a suitable measure of the extent to which the framed-agent is *exploited* in equilibrium.

Specifically, as $D(f_P||P)$ increases, the equilibrium probability that she is in cognitive state $H$ increases. This is due to the fact that the amount the individual loses by purchasing contract $C_F$ is also increasing in this divergence measure. To see this,

we can compute this utility of purchasing contract $C_F$ under the true beliefs, $P$ to be given by

$$U_H(C_F, f) = -\frac{\theta^2}{c} D(f_P||P)$$

which is obviously decreasing in $D(f_P||P)$. More distortion leads to greater utility losses which, due to the fact that cognition is endogenous, leads to a higher cognitive strategy. This also implies that, holding the marginal cost of cognition fixed, the cognitive strategy of the agent is a suitable measure of the extent to which the agent is being exploited by the principal.

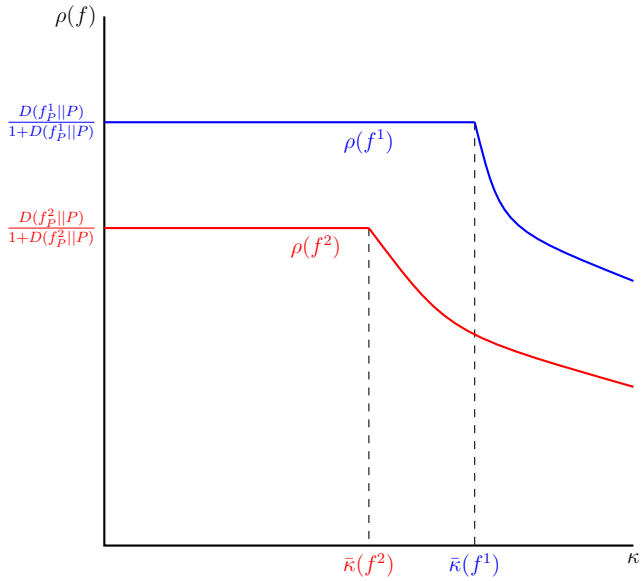


Figure 2.3: Equilibrium Probability of Exiting the Frame, $\rho(f)$

**Example 2.2.** *Recall Example 1 where the set of belief distortions was induced by the use of example-making. The divergence measure for an example-frame was computed to be*

$$D(f_P^E||P) = \frac{1}{P(E)} - 1.$$

*As we previously stated, this implies that the more 'unexpected' the set of examples used (or the less representative the examples were of the entire state space) the greater was the amount of distortion. Proposition 2.1 implies that the more 'unexpected' the set of examples, the greater will be the agent's incentives to invest in cognition.*

*This relates to evidence in the psychology literature that suggests individuals are more likely to engage in causal reasoning (ask questions like 'why did this occur?') the more unexpected an event is (Clary and Tesser (1983), Hastie (1984)). This example suggests that this may also carry over into the domain of framing, where the more "unexpected" the frame is, the more likely the individual is to ask 'why is the contract being framed in this way' and avoid the belief distortion.*

We now derive a functional form for the expected profits of the principal in an arbitrary $f$-frame equilibrium. This expectation will be taken with respect to both uncertainty in $\Omega$ (since the principal holds the true beliefs, $P$) and the cognitive strategy of the agent (since the principal can not observe her cognitive state directly). This computation is provided in the following lemma.

**Proposition 2.2.** *The equilibrium profit of the principal, $\pi(f)$, as a function of frame $f$ is as follows:*

$$\pi(f) = \begin{cases} \frac{\theta^2}{2c} & \text{if } \kappa \leq \bar{\kappa}(f) \\ (1-\rho(f))\frac{\theta^2}{2c}(1+D(f_P||P)) & \text{if } \kappa > \bar{\kappa}(f) \end{cases}$$

*The profit on each individual contract sale is strictly increasing in $D(f_P||P)$ and $D(f_P||P)$ is a sufficient statistic for $f$ in the frame-optimization problem of the principal.*

*Proof.* See Appendix. ■

Combining Proposition 2.1 and Proposition 2.2 illustrates the fundamental trade-off the principal faces when deciding on how to frame the contract. From Proposition 2.2, we get that the principal is able to exploit the framed-type more as he increases the extent to which beliefs are distorted (increases $D(f_P||P)$). Proposition 2.1, however, shows that the the greater the divergence between the frame and the prior, the greater the incentives for the agent to invest in cognition and escape the frame provided by the principal. It follows that the higher is the per-unit profit on the contract, the higher is the probability the agent invests in cognition. Thus, the trade-off of the principal can be summarized as follows: increase the per-unit profit on a contract sale by distorting the agent's beliefs further from the prior at the cost of a lower proportion of exploitable agents, and hence less contract sales.

It should be reinforced that this trade-off is only present due to the fact that the cognitive strategy of the agent has been endogenized. Indeed, if the cognitive-state of the agent was determined by some constant, exogenous probability, then $1 - \rho(f)$ would be constant in $f$ and the principal would find it optimal to use the maximally exploitative frame in order to realize maximal profits. Thus, endogenous cognition is a strongly moderates a principal's ability to engage in exploitation.

### 2.4.2 Optimal Frame

Proposition 2.2 also shows that the choice of optimal frame can be simplified to choosing a value of the $\chi^2$-divergence measure of belief distortion. Let $\bar{D} = \frac{1}{\min_{\omega} P(\omega)} - 1$ which is the maximal value that $D(f_P||P)$ can take. It is obvious that the range of

$D(\cdot||P)$ as a function of $f$ (written $D(\mathcal{F}||P)$) must lie in the set $[0, \bar{D}]$[3]. The frame-optimization problem of the principal can then be re-written with a change of variables as follows:

$$\max_{d \in D(\mathcal{F}||P)} (1 - \rho(d)) \frac{\theta^2}{c} (1 + d)$$

where $\rho(d)$ is given in Proposition 2.1 by replacing $D(f||P)$ with $d$. The solution to this optimization problem in the case where the principal is unconstrained in his ability to distort a framed-type's beliefs is provided in the following proposition.

**Proposition 2.3.** *Let $\mathcal{F} = \Delta(\Omega)$ and define $\underline{\kappa} \equiv \theta^2/c$ and $\bar{\kappa} \equiv \frac{\theta^2}{c}(1 + 2\bar{D})$. The optimal frame $f^*$, as a function of the model's parameters is such that:*

*(a) If $\kappa \leq \underline{\kappa}$, then any $f_P \in \mathcal{F}$ is optimal.*

*(b) If $\kappa \in (\underline{\kappa}, \bar{\kappa})$, then any $f$ such that*

$$D(f_P||P) = \frac{1}{2}\left(\frac{\kappa c}{\theta^2} - 1\right) \tag{2.6}$$

*is optimal.*

*(c) If $\kappa \geq \bar{\kappa}$, then the optimal frame has $D(f_P||P) = \bar{D}$.*

*Proof.* See Appendix. ■

A plot of the optimal frame as a function of $\kappa$ is provided in Figure 3. As can be seen, as the marginal cost of cognition increases, the principal chooses a more exploitative frame, and as the ratio $\theta^2/c$ increases (which captures the amount of surplus

[3]Note that $D(P||P) = 0$, $D(f||P) > 0$ for all $f \neq P$ and $D(f||P) \leq \bar{D}$ for all $f \in \mathcal{F}$. Since $D(\cdot||P)$ is continuous in $f$ and has domain in a convex set, it follows that the range of $D(\cdot||P)$ must be connected. Hence, there must exist an $f \in \mathcal{F}$ such that $D(f||P) = d$ for all $d \in [0, \bar{D}]$.

that can be created through trade), the principal opts to use a less exploitative frame.

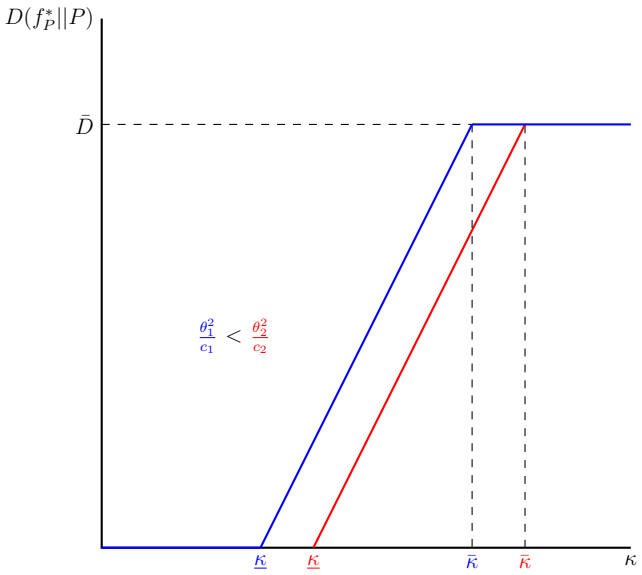


Figure 2.4: Optimal Frame as a Function of $\kappa$

In order to provide intuition for the result, we will impose the following equilibrium refinement: if no distortion is optimal ($P \in \arg\max \pi(f_P)$), then the principal selects $f_P = P$. This equilibrium refinement is such that (a) total surplus subject to the equilibrium conditions is maximized, and (b) the $\chi^2$-divergence of the optimal frame is continuous in $\kappa$. Here, it implies that $f_P = P$ for $\kappa \leq \underline{\kappa}$.

When $\kappa$ is sufficiently small, then in all $f$-frame equilibria, the principal must be indifferent between offering $C_F$ and $C_R$. Given, this the aggregate profit that the principal can earn is constant across all states, given by $\theta^2/(2c)$ (Proposition 2.2).

For intermediate levels of $\kappa$, there now exist $f$-frame equilibria in which the principal finds it optimal to provide only the framed-optimal contract and earn strictly higher profits than the rational benchmark. This, of course, needs to be balanced against decreasing the fraction of the population that is willing to purchase $C_F^f$. This

trade-off is resolved by setting the divergence measure equal to the expression given in (2.6).

The reason that this expression is increasing in $\kappa$ is due to the fact that, as the marginal cost of cognition increases, both $\rho(f)$ and the rate that $\rho(f)$ responds to increased distortions decreases. This incentivizes the principal to choose a more exploitative frame. Similarly, this expression is decreasing in $\theta^2/c$ because an increase in this ratio implies the agent has 'more to lose' by not investing in cognition, making $\rho(f)$ more sensitive to $D(f_P||P)$, thus decreasing the principal's incentives to distort beliefs. It should be noted that both of these effects are due to the fact that the agent's cognitive strategy is responsive to frame choice (and resulting contract offers) of the principal, displaying again the importance for allowing the cognitive-state of the agent to be endogenously determined.

Finally, when $\kappa$ is sufficiently large, cognition is overly costly for the agent and the principal optimally utilizes the maximally exploitative frame, which is that which places probability one on the lowest probability event under the prior. This maximizes the principal's per-unit profits on the sale of a contract and, as $\kappa$ increases, only decreases the likelihood the agent is in a high-cognitive state, increasing aggregate profits further.

The choice of the optimal frame leads to many intuitive, but nonetheless interesting, comparative statics. The following proposition describes how changes in the model's parameters affect the equilibrium cognitive strategy of the agent as well as market efficiency, measured by total surplus. Note that the equilibrium selection criterion that was previously described is also imposed.

**Proposition 2.4.** *Comparative statics are as follows:*

*(a)* $\rho(f^*)$ *is*

*(i) Equal to zero for* $\kappa \leq \underline{\kappa}$,

*(ii) Increasing in* $\kappa$ *and decreasing in* $\theta^2/c$ *for* $\kappa \in (\underline{\kappa}, \bar{\kappa})$, *and*

*(iii) Decreasing in* $\kappa$ *and increasing in* $\theta^2/c$ *for* $\kappa > \bar{\kappa}$.

*(b) Total Surplus (the sum of profits and ex-ante agent utility) is strictly lower than the efficient level (unless* $\kappa \leq \underline{\kappa}$*), is strictly increasing in* $\theta^2/c$, *and is strictly decreasing in* $\kappa$ *for* $\kappa > \underline{\kappa}$.

*Proof.* See Appendix. ■

A global plot of the impact of $\kappa$ on the agent's cognitive strategy is provided in Figure 4.

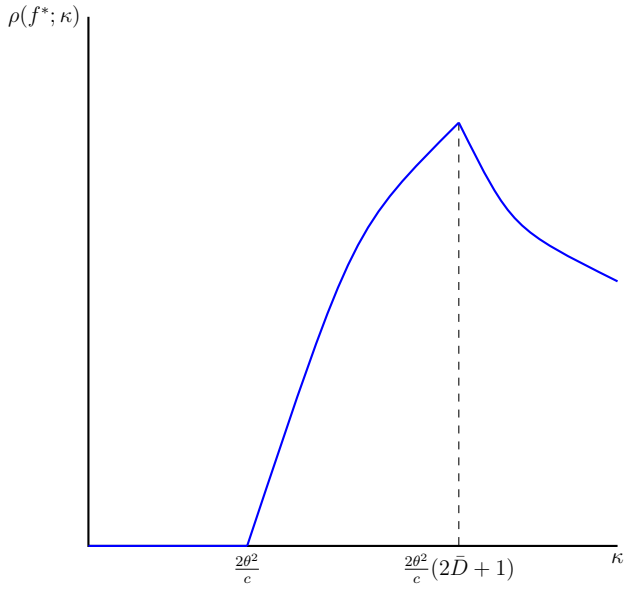


Figure 2.5: Cognitive Strategy Given Optimal Frame, $f^*$

This relationship illustrates precisely why endogenizing the cognition of the agent is of interest. As the marginal cost of cognition increases, there is a direct effect

which depresses the extent to which the dis-incentivizes investment in cognition. There is, however, a competing equilibrium effect which increases the benefits of cognitive effort. Indeed, as $\kappa$ increases, the principal finds it optimal to use a more exploitative frame (Proposition 2.3). Since increased distortions lead to greater utility losses (relative to the true beliefs, $P$), the agent will wish to invest more in cognition to avoid this exploitation. Under the functional form assumptions, whenever both forces are present, the latter effect dominates and one should expect greater investment in cognition for types that find cognition more expensive.

It should be noted that, while this prediction is sharp under the functional form assumptions of the model, it is less clear that the equilibrium effect would *always* dominate the direct effect with a more general class of cost functions (or preferences). One can see, however, that in a more general, if there is a level of $\kappa$ such that the principal would be willing to use the frame $f_P = P$, then there must at least be some interval over which cognitive investment is increasing in its own cost. This is because, for any class of preferences or cost functions, with $\kappa$ sufficiently large, the principal (a) will wish to distort beliefs to some extent and (b) the agent will wish to invest in cognition to avoid at least some utility losses. In order to move from the cognitive strategy of zero to some positive level, there must be a region in which cognition increases in its own cost.

Globally, the relationship between $\kappa$ and cognitive investment is non-monotonic: once $\kappa$ becomes sufficiently large, the optimal frame choice of the principal becomes constant (given by the maximally exploitative frame), only the direct effects of cognitive costs on $\rho(f^*)$ matter and $\rho(f^*)$ begins to decrease.

Similarly, an increase in $\theta^2/c$ implies scales the utility loss the agent faces (if framed) which has a direct effect of increasing the incentives to invest in cognition. There is an equilibrium effect, however, by which the principal exploits less as this ratio increases (Propostion 2.3). Again, with these functional form assumptions, the equilibrium effect dominates the direct increase in cognitive investment and the agent *decreases* her cognitive investment as $\theta^2/c$ increases for intermediate levels of $\kappa$.

The proposition also makes important predictions for market efficiency. A plot of total surplus, together with principal profits and the ex-ante welfare of the agent, is provided in Figure 5.

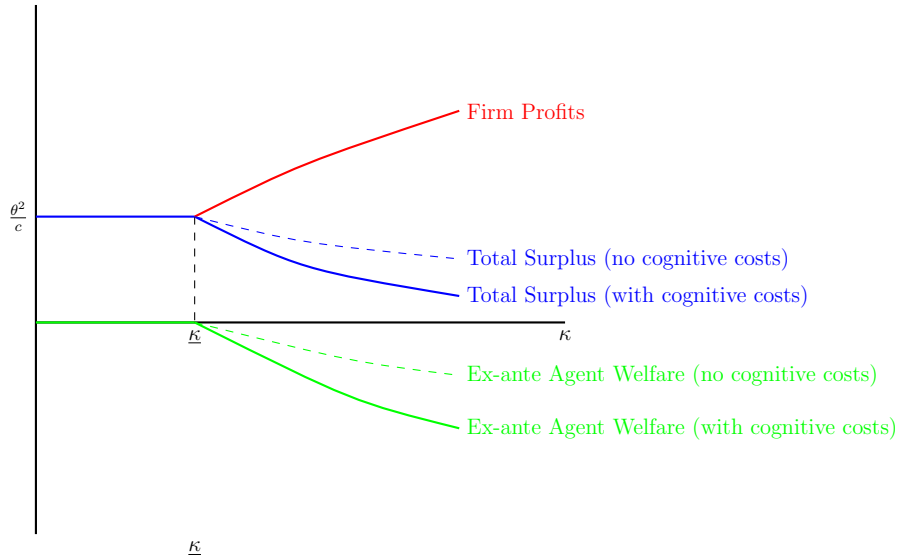


Figure 2.6: Total Surplus as a Function of $\kappa$

This plot shows that, even though the agent is increasing her cognitive strategy in a response to increased exploitation, and thus is not exploited as *often*, the *amount* by which she is exploited more than offsets this. Thus, the ex-ante welfare of the agent decreases in her marginal cost of cognition. Moreover, this more than compensates for increased firm profits and total surplus also falls. Hence, the market is less efficient

when there is a greater concentration of agents with high-cognitive costs. The dashed lines show that this is true irrespective of whether cognitive costs are included in the calculation of agent welfare or not. This implies that market efficiency is not simply falling due to increasing cognitive costs: rather it is decreasing in $\kappa$ due to the principal's increasing exploitation of the agent. While intuitive, the model confirms that a regulator concerned with efficiency should be focused on searching for markets in which the consumer base finds investment in cognition to be overly costly.

## 2.5 Screening

Suppose that, instead of being restricted to offering a single contract, the principal is able to engage in cognition-based screening by offering the agent a menu of contracts. As previously discussed, being forced to choose from a menu itself interact with the cognitive-state of the agent. To shut this effect down, we assume that the presence of a menu to choose from has zero impact on the agent[4]. Additionally, we assume that the principal must apply the same frame, $f_P \in \mathcal{F}$, to all contracts offered in a menu, $M$.

Allowing for the principal to select a menu of contracts usually introduces additional complexity in the form of incentive constraints. This, however, is a non-issue here. More specifically, the optimal contract to offer framed-types, $C_F$, and the rational contract, $C_R$ (both provided in Lemma 2.1) also satisfy their respective incentive constraints. To see this, suppose that both $C_F$ and $C_R$ are offered in the menu under some frame $f$. Obviously the unframed-type will not purchase the framed-types contract (since it has been shown that $U_R(C_F, f) < 0 = U_R(C_R, f)$). Moreover, the framed-type is indifferent between $C_F(f)$ and $C_R(f)$ since the fact that $C_R(f)$ offers

[4]A more general theory may have that there is a probability, $\phi$, such that if a menu of contracts is provided, the agent *exogenously* transitions to a state of conscious decision-making. We take extreme case where $\phi = 0$ in order to maximally contrast the screening and single-contract frameworks.

constant consumption in *all* states of the world implies that types $F$ and $R$ value $C_R$ identically. Hence, if in the low-cognition state, the agent is willing to purchase contract $C_F$. It follows that the optimal menu of contracts for a given $f$ is $M = \{C_R, C_F\}$.

Since the design of the rational contract is such that type $R$ is indifferent between $C_R$ and her outside-option, it follows that the incentives to invest in cognition for a given $f$ are also unchanged from those in the baseline version of the model. The main effect of allowing for a menu of contracts is on the principal's optimal choice of frame, which will impact the cognitive strategy of the agent on the path of play. Comparisons between the equilibrium cognitive strategy of the agent and market efficiency between the screen and no-screen conditions are provided in the following proposition. The equilibrium refinement that if the principal is indifferent between all frames, he chooses that which involves no exploitation is also imposed.

**Proposition 2.5.** *Let $f^S$ be the optimal frame in the screening problem and recall the optimal frame choice in the single-contract frame, $f^*$ (from Proposition 2.3). Then,*

*(a) For all $\kappa < \bar{\kappa}$, $D(f_P^S||P) > D(f_P^*||P)$.*

*(b) For all $\kappa < \bar{\kappa}$, $\rho(f_P^S) > \rho(f_P^*)$.*

*(c) There exists a $\kappa^* \in (\underline{\kappa}, \bar{\kappa})$ such that, when $\kappa < \kappa^*$ total surplus is maximized when screening is not allowed and when $\kappa > \kappa^*$ total surplus is maximized with screening permitted.*

*Proof.* See Appendix. ■

Figure 4 provides a plot comparing the equilibrium cognitive strategy when the principal can screen versus when he can not.

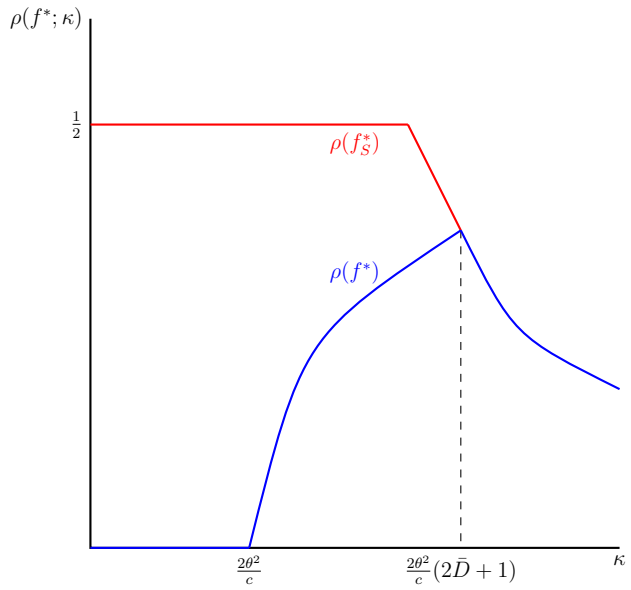


Figure 2.7: Comparing Cognitive Strategies with Screening and Without

As can be seen, the cognitive strategy of the agent is always higher in the screening condition than in the single-contract condition, except when $\kappa$ is sufficiently large. This is because, when able to screen, the principal now makes profit on the rational type, rather than shutting this type out of the market as in the single-contract case. This implies that the principal has an incentive to use a more exploitative frame in equilibrium (increase $D(f_P^S||P)$) and, as such, the probability the agent is in cognitive-state $H$ also increases.

This result displays further the importance of allowing the cognitive strategy of the agent to be endogenously determined. It shows that individual's will be more cognizant of framing effects when faced with a menu of contracts. Moreover, this result is generated without requiring, *a priori*, that the agent responds differently to menus of contracts due to some exogenous process. Instead, it is the result of equilibrium interaction with her environment.

The result also implies that, in terms of market efficiency, neither the screening nor single-contract settings strictly dominate each other. Instead, we find that market efficiency is maximized by banning cognition-based price discrimination if and only if the cost of cognition is sufficiently small.

The intuition for this finding is as follows. When the agent has a low marginal cost of cognition, the principal is willing to use the frame $f_P = P$ in the no-screen condition, and total surplus is maximized. On the other hand, the principal *always* chooses to distort beliefs to some extent when utilizing a menu of contracts. Hence, market efficiency is maximized when cognitive-based price discrimination is not allowed.

In contrast, when $\kappa$ becomes sufficiently large, the principal decides to utilize the maximally exploitative frame, regardless of whether he is permitted to offer a menu of contracts or not. Given that the extent of exploitation is identical and the rational type receives the same utility across settings, it follows that ex-ante consumer welfare is constant. The principal, however, is now earning higher profits by screening as he is able to earn profits on the rational type. Hence, allowing for screening maximizes total surplus.

## 2.6 Conclusion

This chapter has shown that the cognitive-state of an agent is an important moderator of a principal's ability to use framing effects for exploitative purposes. For a specific class of framing effects based on belief distortions, we characterized the optimal frame in terms of a measure of divergence between the frame-induced beliefs and the true beliefs the agent should hold. It is shown that, when the marginal cost

of cognition is small, the principal finds it optimal not to employ framing effects, and the incentives to exploit are increasing in such costs.

The importance of allowing cognition to be determined endogenously has been displayed on a consistent basis. First, endogenous cognition is the key factor that ensures the principal does not find it optimal to distort beliefs maximally. Second, it leads to interesting equilibrium effects, with resulting comparative statics, that would be unattainable in a model where the cognitive-state of the agent was determined according to some exogenous process. In particular, we have shown that the agent's incentives to invest in cognition can actually be increasing in its own cost. This is due to the equilibrium effect of increased exploitation dominating the direct impact of the 'price' increase. Many other comparative statics are derived, all of which are a result of the equilibrium determination of the agent's cognitive state.

We also investigated how allowing for cognition-based price discrimination impacts the model's predictions. It was shown that the principal always chooses a more exploitative frame when able to offer a menu of contracts and the agent is more likely to engage in conscious thought as a response to this. Even though exploitation is greater in the model with screening, market efficiency (measured by total surplus) is improved as long as the marginal cost of cognition is sufficiently high. These findings are important for a regulator (with a certain set of objectives) who is trying to determine whether intervention in a particular market is required.

Regarding contract design, we showed the precise technique the principal uses in order to exploit the agent when she is in a low-cognitive state. Specifically, we showed that if the optimal frame is such that the agent overweights some state (relative to the truth) then the principal will over-provide consumption in this state. Conversely,

if the frame leads the agent to underweight a particular state, then the principal will under-provide consumption in that state. We showed that the principal does this in such a way that the optimal framed contract provides the same level of expected consumption (taken with respect to the true beliefs) as would be offered in the rational benchmark, but the agent's willingness to pay for this level of expected consumption is increased. Hence, the principal is able to increase the price he charges without passing on rents to the agent.

# Chapter 3

# Goal-Setting with Endogenous Awareness

## 3.1 Introduction

Why do individuals repeatedly set goals for themselves that they do not achieve? In the psychology literature, goal-setting has received a great deal of attention (see Locke and Latham (2002) for an extensive review of these studies). Based on the evidence in both the psychology and economics literatures, it is clear that there are three main aspects of goal-setting that should interest economists:

(1) People are motivated to set goals;

(2) Goals tend to induce greater effort and, presumably, improve outcomes for individuals; and

(3) While goal-setting is effective, people *consistently* set goals that they end up falling short of. For example, Markle et al. (2015) find that only 25% of a group of surveyed marathon runners achieved their self-set goal for finishing time. In another paper, Allen et al. (2016) make explicit note of this by stating "goals

are clearly related to expectations ... but goals are not rational expectations." Della Vigna and Malmendier (2006) find evidence that a significant fraction of gym attendees end up paying more in monthly fees than they would have paid per visit, which is indirect evidence that some individuals may fall short of their goal to 'get fit'.

The main objective of this paper is to provide a theoretical model that can account for these three empirical facts, while offering novel predictions that can spur new investigation. The key concept introduced in this paper is the notion of *endogenous awareness:* the agent may be aware of different subsets of the states of the world at the *planning stage* (i.e evaluating and setting goals) versus the *action stage* (i.e choosing effort). Specifically, the theory posits that the agent is fully aware of all possible states of the world that might realize when planning but, instead, when one actually realizes and the action is carried out, is aware of only the realized state and a subset of the counterfactual states that *could have* occurred. This awareness set is endogenously determined by the *type* of goal the individual uses and goals need only be achievable for states in this set.

The details of the model are as follows. An individual (she) faces a task in which she must choose an intensity of effort provision. The cost of effort is stochastic and is drawn from a distribution with finite support. Before the realization of costs, the agent selects a (potentially state-contingent) goal for her future self. The individual is assumed to display time-inconsistency in the form of quasi-hyperbolic preferences (Phelps and Pollak (1968), Laibson (1997)) and accurately measures the extent of this self-control problem (i.e. is fully sophisticated). Goals are then helpful to the individual for regulating this self-control problem.

The agent is assumed to have reference-dependent, loss averse preferences, as proposed by Kahneman and Tversky (1979). The self-set goal forms a reference-point for the individual's future self and her loss aversion will ensure that goal-setting can be an effective tool for self-regulation. Specifically, the model is most closely related to the theory of reference-point formation proposed by Kőszegi and Rabin (2006). In their model, the reference point is a distribution of utility values that the agent expects to receive in each state of the world, and the individual compares the utility of taking a particular action to what she expects to receive in *all* states. The constraint the authors impose is that reference-points should satisfy rational expectations: the utility the agent *actually* receives in each state of the world should be equivalent to what she *expects* to receive.

The theory of endogenous awareness relaxes the requirement that rational expectations must be satisfied in *all* states of the world, which allows for goal deviation. This is operationalized in the paper by assuming that a goal consists of a subset of the state-space (the endogenously determined *awareness set*) and recommended levels of effort for each state in the awareness set. The type of goal utilized by the agent is then determined by the cardinality of the awareness set and rational expectations need only be satisfied for states in this set.

A *fully complete* goal specifies a complete state-contingent plan of action in each cost state that may occur and, in turn, makes the doer fully aware of all these potential states. Given this, the doer will perform (or be subject to) a Kőszegi and Rabin (2006) stochastic-reference-point calculation over all cost-states. At the other extreme, a *fully incomplete* goal is an unconditional statement to one's future self that recommends a single level of effort. It is assumed that such goals make the doer focus on only a single cost-state that is representative (in her mind) of the

entire state-space. Reference-point calculations will then only incorporate utility comparisons to, or within, this representative state, while other *cross-state* utility comparisons are shut-down. This will be shown to be potentially advantageous to the agent: an incomplete goal relaxes the incentive constraints imposed by the doer's actions and induces higher levels of effort.

More generally, the individual also has access to a full range of *partially-complete* goals, which make her future self aware of a chosen set of events that comprise the state-space. This allows the model to address important questions regarding the *optimal degree of state-contingent planning*, and the optimal awareness set that the planning self wishes to induce.

The main takeaway from the model is that the individual, in general, benefits from decreasing the awareness of her future self through use of an incomplete goal. This is due to the fact that the future self of the agent will not compare the utility of effort to counterfactual states of the world that did not occur (as she would with a complete goal). By shutting these comparisons down, an incomplete goal relaxes the incentive constraints imposed by the doer's actions which means that higher level of effort can be induced.

The paper first focuses on a simplified setting in which the state-space is binary: there is a high-cost state and a low-cost state. In this setting, the goal-setter must choose between either a fully complete or a fully incomplete goal. There are four key results. First, I show that an incomplete goal constitutes an optimum for a broad region of the parameter space. In particular, individuals who are either very time-inconsistent or moderately time-inconsistent with intermediate levels of task-dependent ability always find it optimal to utilize incomplete goals. Since incomplete

goals are only coarsely measurable with respect to the state-space however, they are *blunt* regulatory tools. As the individual becomes more time-consistent, the optimal incomplete goal induces an over-provision of effort in the high-cost state that is magnified as the self-control problem dissipates. It follows that complete goals are more effective self-regulatory devices for relatively time-consistent individuals.

Second, the model predicts that the only form of equilibrium goal deviation involves the individual falling short of her effort prescription in the high-cost state. This fits with evidence that goal deviation generally involves an individual falling short of their expectations. A related interpretation of this result is that the individual is *endogenously* exhibiting the *planning fallacy*: an analyst observing the individual would conclude that the agent is systematically under-estimating her ability to achieve her goals. Buehler, Griffen, and Ross (2002) argue that one explanation for the planning fallacy is that people fail to consider all possible outcomes. This paper shows that failing to consider all possible states-of-nature has its advantages and can lead to greater effort provision than would be possible in the counterfactual environment where the agent was forced to use a fully state-contingent plan.

Third, both goal deviation in the high-cost state and the agent's willingness-to-pay for external commitment devices (over the soft commitment provided by goals) display a non-monotonic relationship with the degree of time-inconsistency. This is due to the fact that incomplete goals are very effective self-regulatory devices for individuals with moderate self-control problems: there is an intermediate level of time-inconsistency at which the individual is able to fully motivate her future self and incentivize ex-ante optimal effort. This may provide an explanation for the observation that the take-up of commitment devices is often low, even in the presence

of present bias (Ashraf, Karlan, and Yin (2006), Dupas and Robinson (2013)).

Finally, the model also predicts, somewhat surprisingly, that the degree of self-control for which the magnitude of goal deviation in the high-cost state is maximized, is such that the individual's willingness-to-pay for external commitment is equal to zero (or, equivalently, ex-ante optimal utility is achieved). Thus, an outside actor (parent, teacher, etc) or analyst may *over-estimate* the extent to which paternalistic intervention is necessitated when significant goal deviation is observed. We also derive comparative statics with respect to other key behavioral parameters, including loss aversion and the distribution of costs, which generate additional predictions that can be empirically tested.

We next show that the key results of the baseline version of the model extend to more general, finite state-spaces. In particular, it is shown highly time-inconsistent individuals will generally utilize a goal with some degree of incompleteness. Moreover, I show that, when ex-ante optimal effort is implementable, the planner selects a partition that satisfies a property labeled *extreme-point salience:* in an optimal goal, the doer is only ever made aware of the lowest and highest cost-states. In a further extension, the main results of the baseline model are also shown to be robust to allowing the individual to abstain from goal-setting at the ex-ante stage.

The paper proceeds as follows. Section 2 surveys a selection of the related literature in more detail. Section 3 presents the most general version of the model where the state-space is arbitrary and the individual is assumed to utilize some form of goal-setting. Section 4 provides complete results in the case where the state-space is binary, including a full characterization of the optimal goal and relevant comparative statics. Section 5 analyzes the general model with an arbitrary (but finite)

state-space. Section 6 discusses further extensions of the baseline framework, as well as some of the key modeling assumptions. Section 7 concludes and presents some directions for further research. All proofs are relegated to the Appendix.

## 3.2 Related Literature

The potential links between goal-setting, and reference-dependence and loss aversion (Kahneman and Tversky (1979)) have long been recognized. In an early contribution, Heath, Larrick, and Wu (1999) argue that a goal may form a reference point and, given the presence of loss aversion, an individual will not wish to fall short of this goal. However, their analysis simply assumes the presence of a goal-induced reference point and the authors do not formally model the endogenous determination of self-set goals.

This paper most closely complements a small but growing economics literature on the efficacy of goals to ameliorate self-control problems. Hsiaw (2013) analyzes an optimal stopping-time investment problem with an option value of waiting, in which a present-biased agent has endogenous reference-dependence stemming from a self-set goal. The author finds that, even when there is no loss aversion, goals induce agents to keep investing longer than they would otherwise, ameliorating the self-control problem. Moreover, it is shown that as the relative importance of psychological utility increases, the individual may over-regulate their behavior and invest beyond the first-best stopping time. Suvorov and Van de Ven (2008) analyze a similar moral-hazard environment to this paper. They find that, at long as the degree of present bias is not too large, individuals can overcome their self-control problems by setting appropriate goals. However, both of these papers use the reference-point formulation of Kőszegi and Rabin (2006) and, consequently, goals must satisfy rational expectations: in equilibrium, they are *always* achieved. This

contrasts with this paper, in which a theory of endogenous awareness allows for individuals to deviate from their self-set goals in equilibrium.

A few papers investigate deviation from an (internally or externally) specified actions. Carrillo and Dewatripont (2008) model 'promises', where internal promises can be reinterpreted as self-set goals. These authors take a reduced-form approach, where the incentives to achieve a goal are shaped by an abstract cost function. When this cost function is sufficiently convex, the marginal cost of downward deviation from a promise is small and it is optimal to set goals that are deviated from in equilibrium. The authors provide two distinct micro-foundations for this cost function in the context of an interpersonal interpretation of the game: costs from damaging external reputation and costs from pecuniary punishments. While the authors note that a model of self-reputation (such as Bénabou and Tirole (2004)) would be a suitable micro-foundation for an intrapersonal interpretation of the game, the paper does not formally pursue this avenue and results pertaining to goal deviation are attributed to the curvature of this abstract cost function. In the current model, in contrast, the incentives to set goals and deviate from them are endogenously shaped by standard behavioral parameters, on the measurement of which there is extensive experimental work. Thus, this theory of endogenous awareness provides novel comparative statics with respect to standard behavioral parameters and should therefore be more amenable to empirical validation than a model with an abstract cost function.

Jain (2009) explores theoretically the goal-setting behavior of individuals who are *naive* regarding their future present bias and finds the intuitive result that naiveté permits unrealistic goals to be used in equilibrium. The current paper shows that naiveté is not at all a necessary condition for goal deviation: the individual is fully sophisticated regarding her future self-control issues and goal deviation is the result

of optimal decision-making, rather than a mistake. Moreover, one may expect an individual to eventually 'learn away' an optimistic naiveté bias (Ali (2011)). This seems a particularly pertinent issue, as in many settings individuals are repeatedly settings goals for themselves (e.g. annual New Year's resolutions, professionals setting goals, marathon runners, etc). Therefore, this theory of endogenous awareness may be a more suitable explanation than naiveté for explaining why goal-setting with deviation is a persistent phenomena across time, within an individual.

Koch and Nafziger (2011) analyze a model in which an individual is restricted to setting a goal that is an expected level of output, which will not be realized until *after* an effort decision is made. This allows for goal deviation, although this stems from exogenous uncertainty over how effort translates into output. Given this, the individual is *never* able to set a goal that is always exactly achieved. In contrast, this paper always allows for realistic goal-setting, since complete goals are always feasible and must satisfy rational expectations. Moreover, in their paper, the *direction* of goal deviation is determined by exogenous uncertainty whereas, in this paper, the individual chooses to set goals that he or she *systematically* falls short of: a common observation in the data.

Regarding the notion of coarse goal-setting introduced in this paper, Hsiaw (2016) analyzes a model in which an individual must complete a multi-stage project. The agent must decide whether to bracket her goal broadly (i.e. in aggregate for the entire project) or narrowly (individual goals for each step). The author finds that narrow goals are useful tools for self-regulation, while broad bracketing may have negative consequences (such as the sunk-cost fallacy). In contrast, this paper considers a different form of 'bracketing' (grouping states of the world, rather than expectations over time) and finds that broad bracketing (in this sense) can actually

be beneficial to a time-inconsistent individual.

The model can also be related to theories of self-regulation that rely on the manipulation of beliefs, such as Carrillo and Mariotti (2000), Benabou and Tirole (2002), and Bénabou and Tirole (2004). In these papers, an individual has an ex-ante incentive to maintain uncertainty, or even distort signals, about the true state of his or her ability in order to induce her future self to exert more effort. In contrast, in this paper, individuals wish to manage the states of the world that their future self is *aware* of. This is an important distinction: incomplete goals are effective because they induce individuals to focus solely a particular state of the world and ignore other *counterfactual* ones that did not realize in assessing how they are performing relative to their goal. Thus, effort incentivization here relies on this *focusing effect*, rather than uncertainty-preserving distortions of beliefs. Moreover, in this model, effective self-regulation operates through the confluence of preferences (reference-dependence and loss aversion) and belief distortion. Well-designed experiments should be able to separate beliefs from preferences and thus, the theory of (un)awareness and more 'belief-based' models of self-regulation should be distinguishable empirically.

## 3.3 The Model

### 3.3.1 Formal Details

There are three time periods, indexed by $t = 0, 1, 2$. A time-inconsistent agent with $(\beta, \delta)$ preferences, where $\beta \in [0, 1]$ and $\delta$ is normalized to one[1], must choose effort $e \in [0, \infty)$ at $t = 1$. For effort level $e$, she incurs immediate cost $C(e, c) = ce^2/2$ in $t = 1$ and receives a delayed benefit $Ve$ in $t = 2$, with $V > 0$. The marginal cost of

[1]This is without loss of generality since, in the model, $\delta$ is not separately identifiable from the benefit factor, $V$.

effort is stochastic, taking values in a finite state-space $\Theta = \{c_1, c_2, \ldots, c_n\}$, where $0 < c_1 < c_2 < \ldots < c_n$. The cost realizes in $t = 1$ *prior* to the choice of effort, but is uncertain at $t = 0$. Let $P$ be the probability measure that captures uncertainty over $\Theta$.

**Material Payoffs:** Label the individual at time $t$, "self-$t$". Since the agent is time-inconsistent, there is a wedge between the level of effort the individual desires at $t = 0$ and what she actually chooses in $t = 1$. Specifically, conditional on cost-state $c \in \Theta$, the optimal level of effort from the perspective of self-$t$ solves

$$\max_e \beta_t V e - c\frac{e^2}{2}$$

where $\beta_0 = 1$ and $\beta_1 = \beta$. The solution to this is given by $e_0^*(c) = V/c$ and $e_1(c) = \beta V/c$, respectively for self-0 and self-1. Hence, if $\beta < 1$, then $e_1(c) < e_0^*(c)$ for all $c \in \Theta$ and self-1 under-provides effort relative to the ex-ante optimum.

**Goals:** Self-0 chooses a goal for her future self, which will form a reference-point for effort provision. A goal is a combination of a subset of the state-space, $\xi \subset \Theta$, and a function $\phi : \xi \to [0, \infty)$, specifying a level of recommended effort for each highlighted state in $\xi$. Denote by $g = (\xi, \phi)$ an arbitrary goal. This formulation allows for the individual to set different types of goals. If $\xi = \Theta$, then a goal consists of a fully state-contingent plan of action. We call a goal of this form *fully complete*. At the other extreme, if $\xi$ is a singleton, then the goal is labeled *fully incomplete*. Finally, self-0 may use a *partially incomplete goal*, which is written on a $\xi \subset \Theta$ with cardinality strictly between one and $|\Theta|$.

Assumption 3.1 specifies the restriction on goals set for highlighted states.

**Assumption 3.1.** *($\xi$-Rational Expectations) Let $e(c, g)$ denote the best-response of self-1 to goal $g$ in cost-state $c \in \Theta$. Then, if $c \in \xi$, it holds that $e(c, g) = \phi(c)$.*

This assumption states that goals should be achieved in the set of highlighted states of the world, $\xi$. We say that there is *goal deviation* if some state $c \notin \xi$ realizes.

**Endogenous Awareness:** A novel contribution of this paper is the idea that self-1's awareness of the state-space can be influenced by the type of goal self-0 utilizes. More specifically, it is assumed that self-1 expects that only states in the highlighted set $\xi \subset \Theta$ will occur in the lead-up to choosing effort provision. This alters her *interim expectations* (her beliefs of what can occur just before the realization of the state), which, in turn, affects the extent to which she is *surprised* by the realization of each cost-state. This is formalized in the following assumption and a timeline of this process is given in Figure 1.

**Assumption 3.2.** *(Endogenous Awareness) Let $g = (\xi, \phi)$ denote an arbitrary goal-choice of self-0. Then, the interim beliefs of self-1 are such that*

$$P(c|\xi) = \begin{cases} \frac{P(c)}{\sum_{c' \in \xi} P(c')} & \text{if } c \in \xi \\ 0 & \text{if } c \notin \xi. \end{cases}$$

Hence, self-1 *conditions* on the subset $\xi$ when forming interim expectations. Due to the assumed structure of self-1's psychological utility, manipulation of the awareness set will influence the actual choice of effort selected by self-1.

**Psychological Utility:** The psychological "technology" through which goals are effective regulators of behavior is reference-dependence and loss aversion. For a

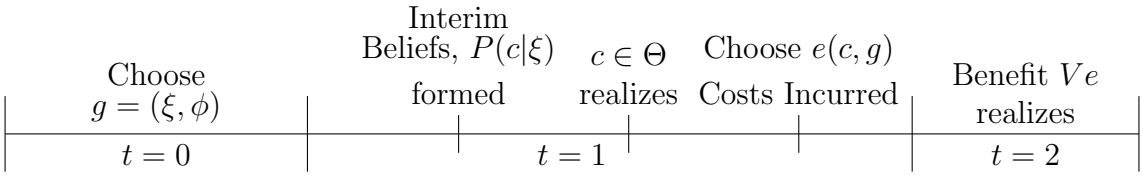


Figure 3.1: Timeline of the Model

given goal $g = (\xi, \phi)$, it is assumed that the individual holds two reference-points: one related to the benefits of effort and a second related to costs of effort. Second, and importantly, it is assumed that the individual compares the costs and benefits of effort not only to the reference points for the state-of-the-world that realizes (if the realized cost-state is in $\xi$), but rather to those for *all* cost-states that are highlighted (that is, all cost-states in $\xi$).

Formally, let $u_1(e, c) = \beta V e$ and $u_2(e, c) = -ce^2/2$. Then, reference-utility around reference-point $i \in \{1, 2\}$ when self-1 chooses effort $e$ in state $c$ relative to what the goal specified in some state $c' \in \xi$ is defined to be

$$\Lambda(u_i(e,c), u_i(\phi(c'), c')) = \begin{cases} \eta(u_i(e,c) - u_i(\phi(c'), c')) & \text{if } u_i(e,c) \geq u_i(\phi(c'), c') \\ -\eta\lambda(u_i(\phi(c'), c') - u_i(e,c)) & \text{if } u_i(e,c) < u_i(\phi(c'), c') \end{cases}$$

where $\lambda > 1$ is the loss aversion parameter and $\eta > 0$ is the marginal rate of substitution between psychological and material payoffs.

**The Maximization Problem:** Self-1 utility from exerting effort $e$ in state $c$ given goal $g = (\xi, \phi)$ is given by

$$U_1(e|c, g) = \beta V e - c\frac{e^2}{2} + \sum_{i=1}^{2} \sum_{c' \in \xi} \Lambda(u_i(e,c), u_i(\phi(c'), c'))P(c|\xi) \tag{3.1}$$

where her psychological utility is the average utility across all states she was aware of when forming her interim expectations. These interim beliefs are given in Assumption 3.2 as a function of $\xi$.

Given this, the optimization problem of self-0 is as follows. Let $e(c, g)$ denote the level of self-1 effort induced in cost state $c$ by the goal $g = (\xi, \phi)$. Then, self-0 chooses $g$ to solve

$$\max_{g=(\xi,\phi)} \sum_{c\in\Theta} P(c) \left( Ve(c, g) - c\frac{e(c, g)^2}{2} \right) \tag{3.2}$$

subject to

$$e(c, g) \in \arg\max_{e} U_1(e|c, g) \quad \text{for all } c \in \Theta \tag{3.3}$$

$$e(c, g) = \phi(c) \quad \text{if } c \in \xi. \tag{3.4}$$

The first constraint is incentive compatibility: self-0 correctly anticipates that self-1 will be best-responding to her choice of goal. The second constraint is Assumption 3.1, which requires that the goal should satisfy rational expectations in all highlighted cost-states.

Notice from (3.2) that the problem of self-0 is to set a goal that induces effort provision as close to the ex-ante desired level as possible. In particular, it is implicitly assumed that self-0 does not incorporate any psychology (dis)utility that self-1 may incur. This assumption is made in order to focus on the role of goals in overcoming the *motivation problem* faced by the individual and to abstract away from scenarios in which self-0 may wish to use goals to also manage her future psychological utility.

A final technical assumption is also made:

**Assumption 3.3.** *The model is analyzed in the region of the parameter space with*

$$\left(\frac{1+\eta\lambda}{1+\eta}\right) < \sqrt{\frac{c_j}{c_{j-1}}}$$

*for all* $j = 2, \ldots, n$.

This assumption will greatly reduce the set of feasible goals that one must search over for an optimum. This assumption is satisfied as long as either $\lambda$ (the degree of loss aversion) is not too large, or each element of the set of cost ratios $\{c_j/c_{j-1}\}_{j=1}^{n}$ is sufficiently large (there is large variation in the potential difficulty of the task).

### 3.3.2 Discussion of Assumptions

We now discuss in more detail and provide justification for the main assumptions of the model. In particular, we relate the notion of $\xi$-rational expectations to the literature, provide a behavioral foundation for endogenous awareness, justify the definition of goal deviation, and the discuss the assumption that self-0 does not incur psychological (dis)utility.

**$\xi$-Rational Expectations:** The concept of $\xi$-rational expectations is a relaxation of requirement in Kőszegi and Rabin (2006) that reference-points should satisfy rational expectations in *all* states of the world. Instead, in this model, rational expectations is only imposed on the set of states self-1 is aware of at the interim stage, $\xi$. Hence, the model is a parsimonious deviation from an established theory in the literature, and the notion of personal equilibrium (defined in Kőszegi and Rabin (2006)) serves as a benchmark that is in the domain of choice of the agent (she can select $\xi = \Theta$). It will be shown that the planner wants to select some $\xi \subset \Theta$. This is due to some odd implications of stochastic reference-point calculations and

illustrates that the cognitive strategy of an individual may involving avoiding such comparisons.

**Endogenous Awareness:** The notion of endogenous awareness is that the set of states self-1 considers possible, just prior to effort choice, is able to be manipulated by self-0. This can be justified through the following behavioral procedure. In order to set a quantitative goal for a state, $c$, the agent must *visualize* herself actually completing the task in that particular cost-state. This process of visualization leads to agent to focus only on the set of pictured states, and, as result, only these highlighted states are conditioned on in forming interim beliefs.

This is equivalent to the idea that self-1 mistakenly believes that the set $\xi$ contains *information* (she says to herself "the only states that can possibly occur are those that I have explicitly planned for") even though in reality there is no information contained in the set and the direction of causality is actually reversed (the set of states explicitly planned actually *become* the only states self-1 believes can occur).

**The Definition of Goal Deviation:** In the model, goal deviation is defined to occur whenever a state $c \notin \xi$ realizes. Given this definition, only when $\xi = \Theta$ will the individual never deviate from her self-set goals with ex-ante probability strictly greater than zero. It will be shown that this is a reasonable definition of goal deviation: in equilibrium, for an arbitrary goal $g = (\xi, \phi)$, when a state $c \notin \xi$ realizes, there will be no value of $\{\phi(c)\}_{c \in \xi}$ that self-1 wishes to choose. Hence, she is deviating from her sub-goals in *every* state that was explicitly planned for, making this a meaningful definition of goal deviation.

It is the case, however, that there is no obvious relationship between any individual cost-state $c \notin \xi$ and the set of cost-states in $\xi$. This poses a conceptual issue for how to quantitatively measure the extent of goal deviation. First, we can measure the goal deviation for a state $c \notin \xi$ from $c' \in \xi$ by computing $D(c, c'|g) = e(c, g) - \phi(c')$. Second, to compute whether the individual is deviating from her goal *on average* in a given state $c \notin \xi$, we can compute the $\sum_{c' \in \xi} D(c, c'|g)P(c'|\xi)$, which is the expectation of the state-by-state measure of goal deviation taken with respect to the agent's interim beliefs.

**No Psychological (Dis)Utility Incurred by Self-0:** This is a simplifying assumption that is made to ensure self-0 utilizes goal to overcome her *motivation problem*, and to abstract away from scenarios in which the individual wants to use goals to also manage her future psychological disutility. Such an assumption would be uncontroversial if self-0 was a separate agent from self-1 (for example, self-0 could be a parent/teacher and self-1 a child/student). In the intra-personal setting, more empirical work is needed to establish the extent to which individuals skew their actions today in order to manage future psychological utility (in addition the effects on material payoffs).

It should be noted that allowing for self-0 to consider the impact of goal-setting on the psychological utility of self-1 would not fully ameliorate the advantages of incomplete goal-setting. Rather, another trade-off (between inducing higher effort at higher psychological cost) is introduced, and the optimal goal would balance this additional trade-off.

## 3.4 Case of a Binary State-Space

In this section, the optimal goal-setting problem of self-0 will be fully characterized in the case in which there are only two cost-states. Denote the state-space by $\Theta = \{c_L, c_H\}$ where $c_H > c_L$ and define $q \equiv P(c_L)$. Here, the individual has a choice between only two types of goals: either a complete goal written on $\Theta$ or a fully incomplete goal written on either $\{c_L\}$ or $\{c_H\}$. For notational convenience, denote the complete goal $g = (\{\{c_L\}, \{c_H\}\}, \{c_L, c_H\}, \phi)$ by $g_C = (g_L, g_H) \equiv (\phi(\{c_L\}), \phi(\{c_H\}))$ and let $g_I^{\xi} \equiv \phi$ denote the incomplete goal $g = (\xi, \phi)$ where $\xi$ is a singleton.

First, the set of feasible complete goals is characterized and the optimal such goal is solved for, in each region of the parameter space. Second, the levels of effort induced by an arbitrary *incomplete goal* are derived and the optimal such goal is characterized. Third, optimal incomplete and complete goals are compared, so as to find the composition of goals that solves self-0's optimization problem. Finally, equilibrium goal deviation, the willingness-to-pay for external commitment, and other comparative statics are analyzed.

### 3.4.1 Complete Goals

We first use the requirement of rational expectations to solve for the set of feasible complete goals, namely the set of $g_C = (g_L, g_H)$ such that

$$U_1(g_L|c_L, g_C) \geq U_1(e|c_L, g_C) \quad \text{for all } e \geq 0$$
$$U_1(g_H|c_H, g_C) \geq U_1(e|c_H, g_C) \quad \text{for all } e \geq 0.$$

Define the *effective threat function*, $\mu$, to be

$$\mu(x) \equiv \frac{1 + x\eta\lambda + (1-x)\eta}{1 + x\eta + (1-x)\eta\lambda}.$$

Notice that $\mu(\cdot)$ is strictly increasing in $x$, $\mu(0) \in (0,1)$, and $\mu(1/2) = 1$. The variable $x$ will be the probability mass of states in which self-1 perceives she is at risk of falling short of her goals and, when filtered through $\mu(\cdot)$, will constitute the effective threat of goals. Indeed, $\mu(1)$ will an upper bound on the degree to which self-0 is able to scale up the desired effort level of self-1 in state $c$, $\beta V/c$. Hence, the function $\mu(\cdot)$ captures how effective reference dependence and loss aversion is for regulating the behavior of self-1. Indeed, when either $\eta = 0$ or $\lambda = 1$ (that is, the individual does not exhibit either reference dependence or loss aversion), then $\mu(x) = 1$ for all $x \in [0,1]$ and goal-setting will be an ineffective tool. Proposition 3.1 describes the set of feasible complete goals as, or equivalently, the set of personal equilibria (Kőszegi and Rabin (2006)) as a function of $\mu(\cdot)$.

**Proposition 3.1.** *In any complete-goal equilibrium, the levels of effort that can be induced in state $j = L, H$ form an interval which is a function of self-1 optimal effort ($\beta V/c_j$) and the effective threat function, $\mu(\cdot)$. Specifically, the set of complete goal equilibria is any $(g_L, g_H)$ such that*

$$g_L \in \left[\underline{g}_L, \bar{g}_L\right] \equiv \left[\mu(0)\frac{\beta V}{c_L}, \mu(q)\frac{\beta V}{c_L}\right] \tag{3.5}$$

$$g_H \in \left[\underline{g}_H, \bar{g}_H\right] \equiv \left[\mu(q)\frac{\beta V}{c_H}, \mu(1)\frac{\beta V}{c_H}\right] \tag{3.6}$$

*Ex-ante optimal effort satisfies these constraints if and only if $q \geq 1/2$ and $\beta = \mu(1-q)$.*

*Proof.* See Appendix. ■

A geometric representation of this set is provided in Figure 2. The fact that all feasible complete goals lie above the line $g_L = g_H\sqrt{c_H/c_L}$ means that, in any complete-goal equilibrium, the goal set in the low-cost state must be both strictly

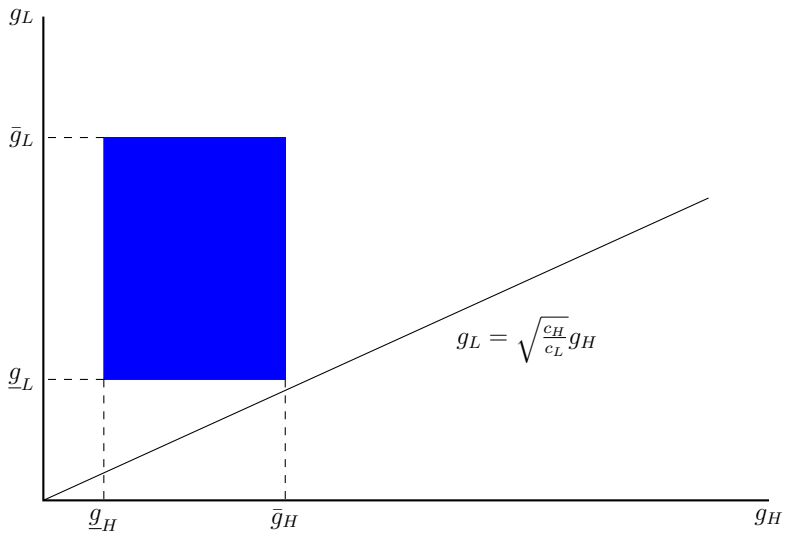


Figure 3.2: Complete Goal Equilibria

higher and strictly more costly to achieve than the goal set in the high-cost state. Assumption 3.3 plays a crucial role in ensuring there are no incentive-compatible complete goals with $g_L \leq g_H \sqrt{c_H/c_L}$, which greatly reduces the space over which one must search for an optimal complete goal.

Proposition 3.1 provides some important implications for the usefulness of complete goals as a self-regulation device. First, notice that self-0's desired effort provision $(V/c_L, V/c_H)$ never satisfies both constraints, except in the special case in which $q \geq 1/2$ and $\beta = \mu(1-q)$. In general, either the maximal level of effort that can be induced in the low-cost state is strictly lower than the optimal level of in this state ($\bar{g}_L < V/c_L$), or the minimum level of effort that can be induced in the high-cost state is strictly higher than the optimal level in this state ($\underline{g}_H > V/c_H$).

To illustrate this, consider the special case in which both cost states are equiprobable ($q = 1/2$). Then, Proposition 3.1 gives $\bar{g}_L = \beta V/c_L < V/c_L$ for all $\beta < 1$, such that the individual can never implement ex-ante optimal effort in cost state $c = c_L$. The reason that no higher level of effort can be induced in the low-cost state

is precisely because of the stochastic reference-point. The threat of falling short of goals is the mechanism that induces higher effort. An increase in $g_L$ induces the individual to implement higher effort in the low-cost state in order to avoid *within*-state psychological disutility. However, this threat is *tempered* by her *cross*-state utility comparison, as she is always outperforming the goal set for the high-cost state (and is incurring a higher cost in terms of utility). This cross-state utility comparison is costly to self-0 and depresses the level of effort that can be implemented in the low-cost state. For the marathon runner, this means that her incentives to exert effort under temperate weather conditions (a low-cost state) are adversely affected by the fact that she would have ran slower and incurred lower costs if the temperature *had of been* abnormally hot (a high-cost state). For $g_L > \beta V/c_L$, the latter effect dominates and thus, self-0 is unable to induce such effort levels in the low-cost state.

This insight is in fact much more general than the two-state case, and also illustrates an 'odd' aspect of stochastic reference-point calculations: not only do they affect the utility that the individual receives, but they also influence actual behavior in ways that are somewhat counterintuitive. This is best illustrated by the fact that, even when the individual is fully time-consistent ($\beta = 1$), only in the very special case in which $q = 1/2$ is it possible for self-0 to induce the first-best effort levels by using a complete goal. For $q < 1/2$, this is due to the fact that there is an under-provision of effort in the low cost state (as outlined above). For $q > 1/2$, this is instead due to an over-provision of effort in the high-cost state (as $V/c_H < \underline{g}_H$ when $q > 1/2$ and $\beta = 1$). Intuitively, this is due to the fact that, under the model's assumptions, in all complete-goal equilibria it holds that $g_L > g_H$. When $q > 1/2$, the low-cost state carries relatively more weight and self-1 increases effort provision in the high-cost state to avoid costly cross-state utility comparisons. This results in

*all* levels of implementable effort in the high-cost state being too high.

With the set of feasible complete goals characterized, Proposition 3.2 provides the solution to the optimal complete-goal setting problem faced by self-0.

**Proposition 3.2.** *The optimal complete goal is unique and is characterized as follows:*

*(a) If $\beta < \mu(0)$, then the optimal complete goal is $g_C^* = (\bar{g}_L, \bar{g}_H)$. There is an under-provision of effort (relative to the ex-ante optimum) in both cost states.*

*(b) If $\beta \geq \mu(0)$, then there exists a threshold $\bar{q}(\beta) \in [1/2, 1]$ such that*

*(i) If $q \leq \bar{q}(\beta)$, the optimal complete goal is $g_C^* = (\bar{g}_L, V/c_H)$. There is an under-provision of effort (relative to the ex-ante optimum) in the low-cost state, except when $q = \bar{q}(\beta)$.*

*(ii) If $q > \bar{q}(\beta)$, the optimal complete goal is $g_C^* = (V/c_L, \underline{g}_H)$. There is an over-provision of effort (relative to the ex-ante optimum) in the high-cost state.*

*Proof.* See Appendix. ■

Figure 3 provides a graphical depiction of the optimal complete goal in the $(q, \beta)$-space, where the exact shape of the solution is given by the fact that $\bar{q}(1) = 1/2$, $\bar{q}(\mu(0)) = 1$, and $\bar{q}(\cdot)$ is strictly decreasing in $\beta$. It shows that, when $\beta$ becomes sufficiently large, ex-ante optimal effort will be induced in the high-cost state if $q$ is sufficiently small, while ex-ante optimal effort for the low-cost state is induced when $q$ is sufficiently large. The intuition for this proposition is as follows. First, when the individual is highly time-inconsistent ($\beta < \mu(0)$), the maximal effort that self-0 can implement with a complete goal is strictly lower than the ex-ante optimum for each

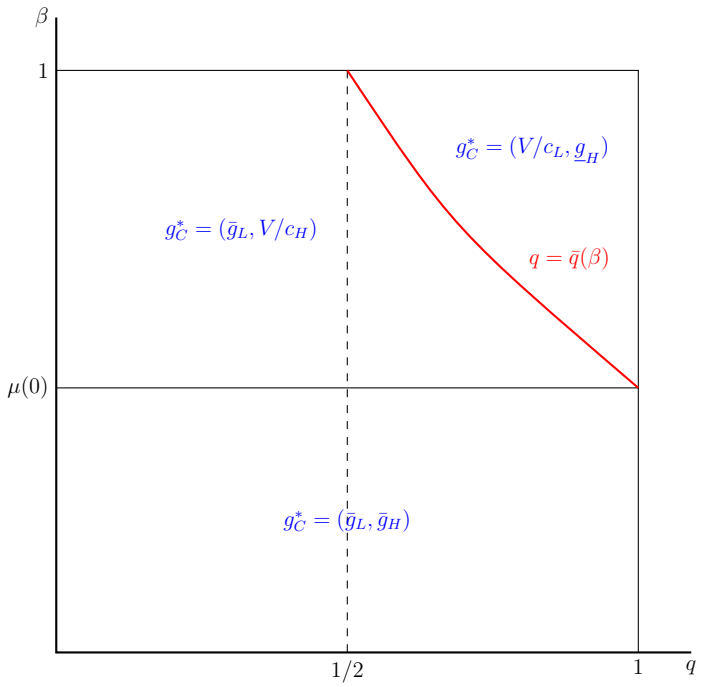


Figure 3.3: Optimal Complete Goals

cost-state. Hence, self-0 chooses to induce as much effort as possible in each state and will set $g_L^* = \bar{g}_L$ and $g_H^* = \bar{g}_H$.

As the individual becomes more time-consistent ($\beta$ increases beyond $\mu(0)$), it is now possible to implement first-best effort in at least one cost-state. When the probability of the high-cost state is relatively large ($q < \bar{q}(\beta)$), it is much easier for the individual to regulate her behavior in the high-cost state. Indeed, self-0 is able to implement ex-ante optimal effort in the high-cost state and will find it optimal to do so. However, with $q$ small, the threat of falling short of her goals is not sufficient for the low-cost state and, as such, all implementable effort levels for the low-cost state are strictly lower than $V/c_L$. Thus, self-0 will choose the maximal low-cost-state goal that is incentive compatible. In contrast, if the low cost-state is sufficiently likely ($q > \bar{q}(\beta)$), it is now possible for self-0 to implement first-best effort in the low-cost state. However, when the probability of the low-cost state is high, the

threat of falling short of goals in the high-cost state is too great, which leads to all implementable efforts for the high-cost state being strictly greater than this state's ex-ante optimum. Hence, self-0 chooses the minimum high-cost-state goal, $\underline{g}_H$, that is incentive compatible.

While complete goals can improve the individual's outcomes, the requirement of rational expectations and full cross-state comparisons has been shown to tighten the incentive constraints faced by self-0 quite drastically in certain regions of the parameter space. For example, when $q < 1/2$, $\bar{g}_L < \beta V/c_L$ for all $\beta \leq 1$ due to the disappointment that state $c_H$ did not realize that self-1 experiences in the low-cost state (as outlined in the previous section). Hence, the individual is doing *strictly worse* in this state than she would in the *absence* of a goal-setting technology. As will be shown in the next section, this is an advantage of incomplete goals: since they induce future unawareness of states that did not materialize, counterintuitive cross-state utility comparisons are not computed by self-1, incentive constraints are relaxed, and higher levels of effort can be implemented.

### 3.4.2 Incomplete Goals

In the model with two cost-states, an incomplete goal can be thought of as a statement of the form "no matter what, exert effort $\phi \geq 0$" coupled with a salient-state that provides a quantitative representation of what self-1 expects to occur. With such a goal, self-1 is aware only of the salient state, $\xi \in \{c_L, c_H\}$, and if the salient state does not realize then this will be perceived as an unexpected surprise. Define the following thresholds:

$$\underline{e}_j \equiv \mu(0)\frac{\beta V}{c_j}, \quad \bar{e}_j \equiv \mu(1)\frac{\beta V}{c_j}$$

for $j = L, H$, where $\mu(\cdot)$ was previously defined. These thresholds will be shown to be, respectively, the lowest and highest levels of effort that self-0 can induce self-1 to undertake in state $j = L, H$ when an incomplete goal is used. The following proposition characterizes the levels of implementable effort for a given incomplete goal $g_I^{\xi}$, as a function of these thresholds.

**Proposition 3.3.** *Suppose that self-0 sets an incomplete goal $g_I^{\xi}$. Then,*

*(a) If $\xi = c_H$, the goal must satisfy*

$$g_I^{c_H} \in [\underline{e}_H, \bar{e}_H] \tag{3.7}$$

*with $e(c_L, g_I^{c_H}) = \underline{e}_L$ and $e(c_H, g_I^{c_H}) = g_I^{c_H}$.*

*(b) If $\xi = c_L$, the goal must satisfy*

$$g_I^{c_L} \in [\underline{e}_L, \bar{e}_L] \tag{3.8}$$

*with $e(c_L, g_I^{c_L}) = g_I^{c_L}$ and $e(c_H, g_I^{c_L}) = \bar{e}_H$. The maximum level of effort in the low-cost state that can be implemented with a $c_L$-salient incomplete goal, $\bar{e}_L$, is strictly greater than that which can be implemented with a complete goal, $\bar{g}_L$.*

*Proof.* See Appendix. ■

As hypothesized in the previous section, the incentive constraints faced by self-0 may be relaxed relative to complete goal-setting. This depends crucially on whether the low-cost state is a feasible candidate for saliency; that is, whether $c_L \in S(\Theta)$. Suppose this is the case and $c_L$ is made salient by self-0. Then, setting $g_I^{c_L} = \bar{e}_L$ leads

to a best response by self-1 in cost-state $c_L$ of

$$e(c_L, g_I^{c_L}) = \mu(1)\frac{\beta V}{c_L} > \mu(q)\frac{\beta V}{c_L} = \bar{g}_L.$$

where the right-hand side of the inequality is the maximum level of effort that can be implemented in the low-cost state with a complete goal. Thus, this incomplete goal allows for greater effort provision in state $c = c_L$ than can be induced with *any* complete goal. This is because the costly cross-state utility comparisons which are built into a stochastic reference-point calculation are suppressed as the incomplete goal maintains self-1's unawareness of the counterfactual, high-cost state when $c_L$ realizes. It follows that self-1 must only consider the within-state psychological costs of not satisfying the reference point, which results in greater effort provision.

To illustrate, recall the example of the marathon runner. With a complete goal, the runner was forced to consider the high-cost state when the low-cost state realized. This was costly due to the fact that a lower level of effort was required under abnormally hot conditions, relative to the temperate-weather state, which depressed the incentives for effort provision. With the incomplete goal, however, if the runner made the temperate weather-state salient, she is now focused solely on the psychological costs and benefits of achieving the task in the realized weather-state. This shuts down self-1's (somewhat unintuitive) disappointment that the temperature wasn't abnormally hot and leads to greater effort provision.

With the set of incomplete-goal implementable effort levels characterized, it is now possible to derive the optimal incomplete goal from the perspective of self-0. This is described in the following proposition.

**Proposition 3.4.** *(a) If $\beta \leq \mu(0)$, then the optimal incomplete goal involves making $c_L$ salient and setting $g_I^{c_L} = \bar{e}_L$. There is an under-provision of effort (relative to the ex-ante optimum) in both cost states, except when $\beta = \mu(0)$.*

*(b) If $\beta > \mu(0)$, then there exists a threshold $\bar{q}_I(\beta) \in (0, 1/2)$ such that*

*(i) If $q \geq \bar{q}_I(\beta)$, then the optimal incomplete goal involves making $c_L$ salient and setting $g_I^{c_L} = V/c_L$. There is an over-provision of effort (relative to the ex-ante optimum) in the high-cost state.*

*(ii) If $q < \bar{q}_I(\beta)$, then the optimal incomplete goal involves making $c_H$ salient and setting $g_I^{c_H} = V/c_H$. There is an under-provision of effort (relative to the ex-ante optimum) in the low-cost state.*

*Proof.* See Appendix. ■

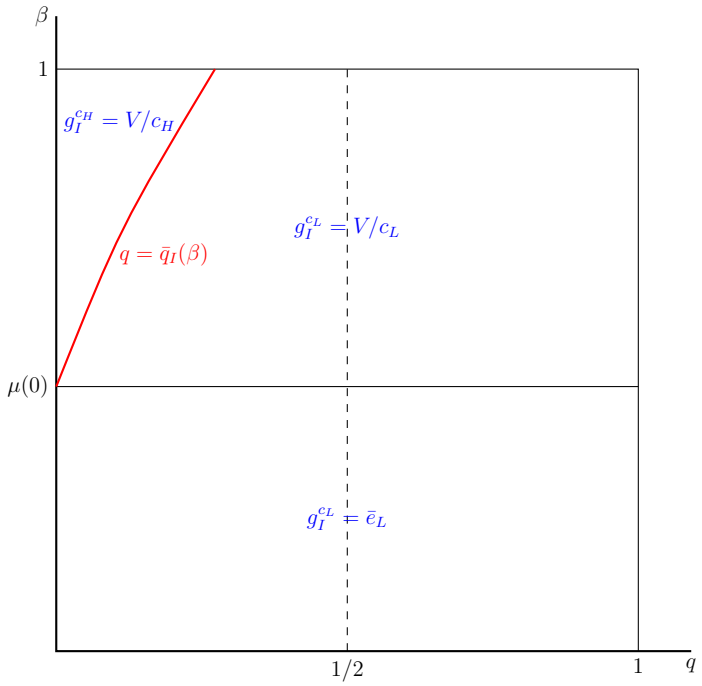


Figure 3.4: Optimal Incomplete Goals

The most important takeaway from Proposition 3.4 is that the optimal incomplete goal always entails a positive ex-ante probability of goal deviation. For $\beta \leq \mu(0)$, the optimal incomplete goal is $g_I^{c_L} = \bar{e}_L$. Since $\bar{e}_H < \bar{e}_L$ under Assumption 3.3, this implies that self-0 is always falling short of the optimal incomplete goal in the high-cost state. For $\beta$ sufficiently large the only incomplete goals utilized are in the set $\{V/c_L, V/c_H\}$. From Proposition 3.3, choosing $g_I^{c_H} = V/c_H$ leads to an under-provision of effort in the low-cost state. However, the minimal level of implementable effort in the low-cost state is still strictly higher than $V/c_H$, so the individual actually *overachieves* relative to her goal in this state. Similarly, choosing $g_I^{c_L} = V/c_L$ induces an over-provision of effort in the high-cost state. The individual still, however, falls short of this goal in the high-cost state since the maximal level of implementable effort in the high-cost state is strictly less than $V/c_L$. Interestingly, since both a $c_H$-salient and a $c_L$-salient incomplete goal can be optimal depending on the difficulty of the task, this implies that, for a person restricted to use incomplete goals as a self-regulation device (perhaps due to a cognitive constraint that limits the individual's ability to engage in state-contingent planning), one could observe both downwards and upwards deviation.

A graphical representation of the optimal incomplete goal for all $(q, \beta) \in [0, 1]^2$ is provided in Figure 4. The intuition is as follows. When the individual is very time-inconsistent ($\beta \leq \mu(0)$), then self-0 desires effort in each state of the world greater than can be induced with any incomplete goal. It follows that the optimal such goal is one that induces $\bar{e}_j$, $j = L, H$, which can achieved by making $c_L$ salient and setting the level of expected effort equal to $\mu(1)\beta V/c_L$.

As the individual becomes more time-consistent, it is now feasible to implement the optimal level of effort in *either* the high-cost or the low-cost state (but not both).

Making $c_H$ salient and setting $g_I^{c_H} = V/c_H$ will induce ex-ante optimal effort in the high-cost state, at the cost of an *under-provision* of effort if the task turns out to be easy. This is due to the fact that $V/c_H$ is a relatively low goal for state $c_L$, which incentivizes self-1 to exert lower effort in this cost state. In contrast, making $c_L$ salient and setting $g_I^{c_L} = V/c_L$ will induce self-0's desired level of effort in the low-cost state, but is a relatively high expectation for the high-cost state, which incentivizes an *over-provision* of effort when the task is difficult. This trade-off faced by self-0 is largely resolved by the relative likelihood of each of the two states. When $q$ is small (and $\beta$ is sufficiently large), optimizing for the high-cost state is more salient and thus, $g_I^{c_H} = V/c_H$ constitutes an optimal incomplete goal. On the other hand, as $q$ grows, the low-cost state has more weight in self-0 expected utility and, consequently, $g_I^{c_L} = V/c_L$ becomes the dominating incomplete goal.

It is also informative to interpret incomplete goal-setting as a form of *visualization* which is a commonly prescribed psychological technique for maintaining motivation to achieve goals. One can think of an incomplete goal with salient-state $c_L$ ($c_H$) as the individual visualizing the "best-possible" ("worst-possible") circumstances that they may need to complete the task under. For example, $\xi = c_L$ is equivalent to the marathon runner visualizing a perfectly temperate day when completing her run. In contrast, $\xi = c_H$ can be thought of as the runner picturing an abnormally hot day and imagining herself exerting effort that optimally matches with these adverse conditions. Under this interpretation, the results of this section suggest that both of these visualization techniques are candidates for optimal behavior. However, it will be shown in the next section that, while making $c_H$ salient can constitute an optimal incomplete goal, it will *never* be an optimum when the domain of choice includes complete goals.

### 3.4.3 Optimal Goals

The previous sections have provided a characterization of optimal complete and incomplete goals. We now explore the trade-offs between complete and incomplete goal-setting. Proposition 3.5 summarizes, and Figure 5 illustrates, the optimal goal-setting behavior for the different regions of the parameter space.

**Proposition 3.5.** *The optimal goals of the agent are as follows:*

*(a) If $\beta \leq \mu(0)$, then the optimal goal is the incomplete goal $g_I^{c_L} = \bar{e}_L$. Goal deviation involves the individual falling short of her goal in the high-cost state.*

*(b) If $\beta > \mu(0)$, then there exists a threshold $\beta^*(q) \in (\mu(0), \mu(1-q))$ such that*

*(i) If $\beta < \beta^*(q)$, the optimal goal is the incomplete goal $g_I^{c_L} = V/c_L$. Goal deviation involves the individual falling short of her goal in the high-cost state.*

*(ii) If $\beta > \beta^*(q)$, the optimal goal is the complete goal given in Proposition 3.2. There is no goal deviation in equilibrium.*

*Proof.* See Appendix. ■

Figure 5 displays the solution to self-0's optimal goal-setting problem. Lemma A.3, provided in the Appendix, solves for some key properties of the threshold $\beta^*(q)$; namely that this function is inverse U-shaped in $q$, with a peak at some point $q^* < 1/2$.

There are a number of important predictions and implications resulting from Proposition 3.5. Since the incomplete goal $g_I^{c_H} = V/c_H$ never constitutes an optimal goal, the model predicts that the individual will never set a goal that she *exceeds*.

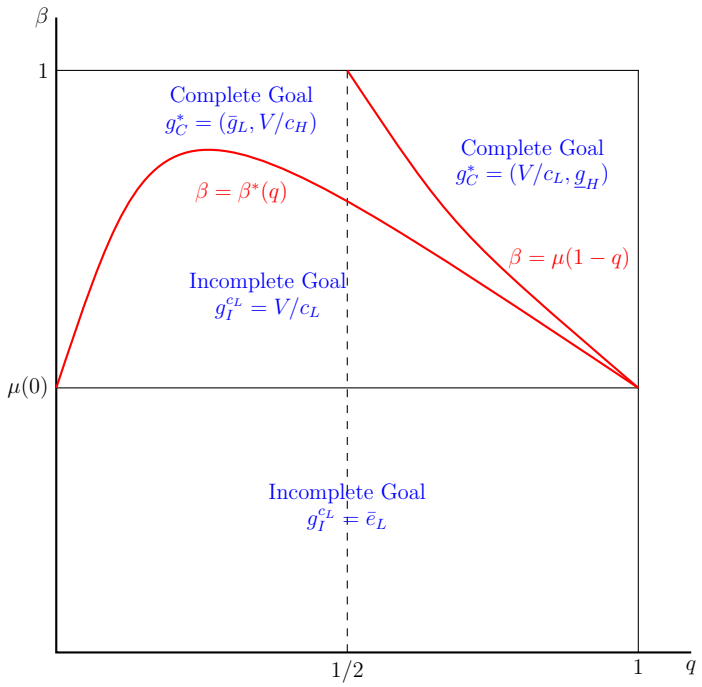


Figure 3.5: Optimal Goals

However, there is a non-trivial region of the parameter space for which the individual is motivated to choose a goal that she knowingly falls short of with positive probability. Thus, the model predicts systematic downwards deviation from goals, which fits well with empirical evidence. Continuing with the reinterpretation of incomplete goal-setting as a form of 'visualization', it suggests that envisioning the "worst-possible" scenario under which a task must be undertaken is never productive, while visualizing an easy task may constitute optimal behavior. This lends credence to a self-help recommendation of the form "visualize tasks being completed under the most desirable conditions".

Proposition 3.5 also provides predictions regarding the types of individuals for which we should expect to observe *any* goal deviation whatsoever. First, one should expect that relatively time-inconsistent agents are more likely to set incomplete goals, while relatively time-consistent individuals should desire the additional flexibility of

complete goals. Second, holding $\beta$ fixed at some intermediate level, one should expect goal deviation from an individual who faces a large amount of uncertainty regarding task difficulty (since $\beta^*(q)$ is inverse U-shaped). Taking a population interpretation of $q$ as a measure of *task-dependent ability*, the model predicts that individuals with intermediate levels of ability are more likely to use incomplete goals and subject themselves to goal deviation. Combining these results, one should expect either *very* time-inconsistent individuals or individuals with moderate levels of both self-control and task-specific capability to set goals that are not always achieved. Empirically, it should be feasible to validate or reject these findings with an appropriately designed experiment.

The intuition for the shape of the solution to the optimal-goals problem is a follows. For low levels of $\beta$, the main objective of self-0 is to select the goal that induces as much effort as possible from self-1 in each cost state. Recall that an incomplete goal with salient-state $c_L$ relaxes the incentive constraints that self-0 faces relative to complete goals, as they do not involve costly cross-state utility comparisons. Hence, a very time-inconsistent individual finds it optimal to use incomplete goals in order to take advantage of these relaxed incentive constraints.

However, an incomplete goal is a blunt instrument for regulating effort provision as it must be constant across states of the world. This is particularly costly for the incomplete goal $g_I^{c_H} = V/c_H$, as it is always strictly dominated by the complete goal $g_C = (\bar{g}_L, V/c_H)$ in the region in which it was optimal. This is because both types of goals induce equivalent levels of effort in the high-cost state, but the additional flexibility afforded by the complete goal allows for self-0 to induce a more desirable level of effort in the low-cost state. For $g_I^{c_L} = V/c_L$, as the individual becomes more time-consistent, self-1 begins to *over-provide* effort in the high cost state as

$\mu(1)\beta V/c_H > V/c_H$ if and only if $\beta > \mu(0)$. Hence, the utility of the incomplete goal $g_I^{c_L} = V/c_L$ decreases in $\beta$ in this region. After $\beta$ increases past the threshold $\beta^*(q)$, this overprovision of effort becomes relatively too costly and the optimal complete goal provides strictly higher ex-ante utility than the optimal incomplete goal.

### 3.4.4 Goal Deviation and Welfare

Now, some further predictions of the model are explored. In particular, since goal deviation is a leading motivation for the development of this theory of endogenous awareness and incomplete goal-setting, it is desirable to explore the comparative statics of some measure of goal deviation. We also show that the ex-ante utility of the individual is maximized at the point in which goal deviation is maximized, which suggests that using the presence of goal deviation to justify paternalistic intervention is misguided.

The previous analysis show that, self-0 never selects a goal which leads to deviation in the low-cost state when complete goals are available as self-regulation devices. Thus, in equilibrium, the individual is not achieving her self-set goals only in the high-cost state. Proposition 3.6 provides a calculation of the magnitude of deviation in this cost-state.

**Proposition 3.6.** *The magnitude by which self-1 falls short of her goal in the high-cost state, $D_H$, is given by*

$$D_H = \begin{cases} \mu(1)\beta V\left(\frac{1}{c_L} - \frac{1}{c_H}\right) & \text{if } \beta \leq \mu(0) \\ \frac{V}{c_L} - \mu(1)\frac{\beta V}{c_H} & \text{if } \beta \in (\mu(0), \beta^*(q)) \\ 0 & \text{if } \beta > \beta^*(q) \end{cases}$$

*It is increasing in $\beta$ for $\beta < \mu(0)$, and decreasing in $\beta$ for $\beta > \mu(0)$.*

*Proof.* See Appendix. ■

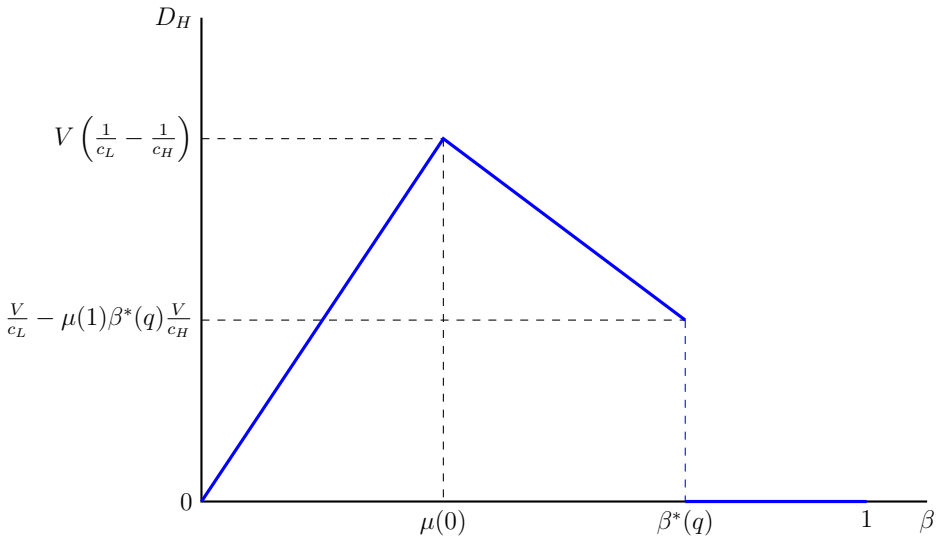


Figure 3.6: Deviation from goal in state $c = c_H$ as a function of $\beta$

Figure 6 displays a plot of goal deviation in the high-cost state as a function of $\beta$. Notice that the degree of goal deviation in cost state $c_H$ is *non-monotonic* in the time-inconsistency parameter. To develop intuition, it is useful to interpret this finding through the lens of the marathon runner. When the runner is very time-inconsistent ($\beta$ is small), she is unable to induce her future self to exert much effort

in *any* possible weather condition, which implies a low level of goal deviation. As $\beta$ increases, however, the runner's optimal goal for effort exertion increases at a rate proportional to $1/c_L$, while her *actual* effort provision in the high-cost state increases at a slower rate (proportional to $1/c_H$). This leads to an increase in the degree to which the runner falls short of her goal when the temperature is abnormally hot. However, when $\beta$ reaches an intermediate threshold, the runner can now achieve her optimal run-time for both weather conditions simultaneously with a well-calibrated incomplete goal. After this threshold, her optimal goal becomes constant in $\beta$ (given by $g_I^{c_L} = V/c_L$) and since increases in $\beta$ continue to increase effort in the abnormally hot weather condition, it follows that the degree of goal deviation decreases. Finally, when $\beta$ reaches a further threshold, there is a discrete drop in goal deviation (to zero) as the marathon runner switches from using an incomplete goal to a complete goal, which is required to satisfy rational expectations and thus, never involves any deviation.

While the non-monotonic shape of goal deviation (in $\beta$) suggests an empirical test to validate the model, it is also important to investigate the resulting impact this has on ex-ante equilibrium welfare. However, since utility is difficult to measure across individuals with heterogeneous degrees of time-inconsistency, any theoretically-predicted impact on welfare is potentially difficult to test. One feasible method for verifying the model's comparative-static predictions on welfare is to use the willingness-to-pay for external commitment.

Recall that the first-best level of utility is attained at $e_0^*(c_j) = \frac{V}{c_j}$ for each $j \in \{L, H\}$, and takes value

$$U_0^{FB} = \frac{1}{2}V^2\left(\frac{q}{c_L} + \frac{1-q}{c_H}\right)$$

which is independent of $\beta$. Define the willingness-to-pay (in $t = 1$[2]) for external commitment (over the soft-commitment device of goal-setting) of self-0 to be given by

$$WTP(\beta) = U_0^{FB} - U_0^*(\beta) = \frac{1}{2}V^2\left(\frac{q}{c_L} + \frac{1-q}{c_H}\right) - U_0^*(\beta).$$

where $U_0^*(\beta)$ is the indirect utility of self-0 (from $t = 1$ onwards) when she only has access to the goal-setting technology. Since $U_0^{FB}$ is independent of $\beta$, it follows that $WTP(\beta)$ adequately captures the model's predictions for equilibrium welfare. The following proposition formalizes how $WTP(\cdot)$ varies with $\beta$.

**Proposition 3.7.** *Suppose that the individual deviates from her goals with positive ex-ante probability ($\beta < \beta^*(q)$). Then, her willingness-to-pay for full commitment as a function of $\beta$ is given by*

$$WTP(\beta) = \begin{cases} \frac{1}{2}V^2\left[\frac{q}{c_L} + \frac{1-q}{c_H}\right](1-\mu(1)\beta)^2 & \text{if } \beta \leq \mu(0) \\ \frac{1}{2}V^2\frac{1-q}{c_H}(1-\mu(1)\beta)^2 & \text{if } \beta > \mu(0) \end{cases}$$

*It is decreasing in $\beta$ for $\beta < \mu(0)$ and increasing in $\beta$ for $\beta \in (\mu(0), \beta^*(q))$.*

*Proof.* See Appendix. ■

Figure 7 provides a plot of $WTP$ against $\beta$. This result has important practical implications for the design and sale of commitment devices. In particular, it implies that, conditional on observing goal deviation, the individual's willingness-to-pay for external commitment is *not* a monotonically decreasing function of $\beta$ (as would be the case if no goal-setting technology were available). This may provide an explanation for the (somewhat) counterintuitive finding that time-inconsistent individuals are

[2]If the individual were to pay for the commitment device in $t = 0$, she would also need to factor in her discount rate of $\beta$. To abstract from her willingness-to-pay being affected *instrumentally* by her discounting, we assume self-0 commits to pay for the device at $t = 1$.

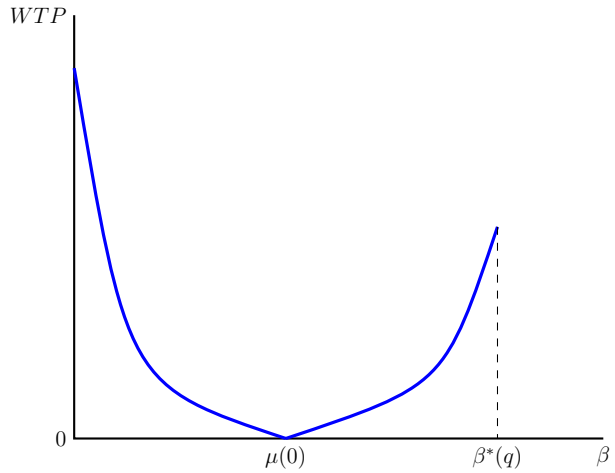


Figure 3.7: Willingness-to-pay for full commitment as a function of $\beta$ with $q = 1/2$

often not willing to purchase commitment devices. Absent an effective goal-setting technology, agents with intermediate levels of time-inconsistency would be willing to pay a reasonable amount for full commitment. However, incomplete goal-setting is extremely effective in this region and, consequently, intermediate $\beta$-types have a low willingness-to-pay. Moreover, individuals with $\beta$ close to $\mu(0)$ are willing to pay less for full commitment than a person who is substantially more time-consistent. Thus, while a researcher or a firm may elicit that an individual is time-inconsistent and thus should demand commitment, actual willingness-to-pay for such commitment may be low due to how effective incomplete goals are for regulating behavior.

The following corollary provides a connection between the magnitude of goal deviation in the high-cost state and the willingness-to-pay for external commitment. Just as Proposition 3.7 shows that observing time-inconsistent behavior is not sufficiently informative to conclude that the individual will demand external commitment, Corollary 3.1 suggests that observing extensive deviation from goals is also not helpful in this regard.

**Corollary 3.1.** *When goal deviation is maximized (at $\beta = \mu(0)$) then, ceteris paribus, the willingness-to-pay for external commitment is equal to zero. Equivalently, the effort choice of self-1 is equal to self-0's desired effort levels when goal deviation is maximal.*

This result immediately follows from Proposition 3.6 and Proposition 3.7 and has important implications for "paternalistic" policymaking. In the self-help literature, there is often a focus on helping individuals learn to set goals that they *can* achieve (for example, the recommendation of S.M.A.R.T. goal-setting, where the "A" stands for attainable). This prescription implicitly assumes that setting achievable goals is the most effective method for improving future outcomes. The finding here suggests that this may not be an optimal approach for motivating people. In particular, as has been seen, inducing individuals to set goals that are achievable in all possible states (formally, that satisfy rational expectations) may lead to costly cross-state utility comparisons that actually restrict the levels of implementable effort. An individual for which incomplete goals constitute optimal behavior may wish to avoid such advice to benefit from relaxed incentive constraints. Indeed, at the point where goal deviation is maximized, the optimal complete goal falls (perhaps, significantly) short of the ex-ante optimum[3].

### 3.4.5 Further Comparative Statics Results

So far there has been a focus on how the degree of time-inconsistency, $\beta$, affects key observables. However, the effects of other behavioral parameters are also of

[3]Of course, this is partly due to the assumption that self-0 does not incorporate any psychological gains and losses into her utility. If she did, then this would likely affect this relationship. However, the extent to which individuals incorporate (and actively manage) future reference utility is, inherently, an empirical question and the assumption is made here for technical convenience.

interest. Moreover, each of the parameters of the model are, arguably, elicitable through experimental techniques. Thus, it should be feasible to empirically test the comparative static predictions of the remaining parameters. As discussed previously, this is an advantage of this model relative to other theories of goal deviation that rely on either an abstract cost function or naiveté, both of which may be more difficult to estimate than the behavioral forces that drive the results here.

Proposition 3.8 presents the effect of marginal changes in the loss aversion parameter ($\lambda$), the cost parameters ($c_L$ and $c_H$), and the benefit ($V$) on the extent of goal deviation, $D_H$.

**Proposition 3.8.** *Suppose that self-0 selects the optimal goal. Then,*

*(a) For $\beta < \mu(0)$, the magnitude of goal deviation in the high-cost state, $D_H$, is increasing in $\lambda$, $V$, and $c_H$, and is decreasing in $c_L$.*

*(b) For $\beta \in (\mu(0), \beta^*(q))$, $D_H$ is decreasing in $\lambda$ and $c_L$, and increasing in $c_H$ and $V$.*

*(c) For $\beta > \beta^*(q)$, $D_H$ is constant with respect to all behavioral parameters.*

*Proof.* See Appendix. ■

The first thing to notice is that the comparative statics with respect to the loss aversion parameter are identical to those with respect to $\beta$: both $\beta$ and $\lambda$ increase goal deviation for low levels of $\beta$, decrease goal deviation for intermediate levels of $\beta$, and decrease the likelihood that the conditions under which equilibrium goal-setting involves deviation will be satisfied. This implies a sense in which $\beta$ and $\lambda$ are

*substitutes* in the management of effort provision[4].

An increase in the cost ratio, $c_H/c_L$, unambiguously leads to an increase in $D_H$ at all levels of time-inconsistency, and makes it more likely that incomplete goals constitute optimal behavior. Thus, goals set for tasks to be completed in a setting with more homogeneous conditions (e.g. athletes performing indoors) should be achieved more often than those in which the conditions are likely to be more variable (e.g. if to be performed outdoors). To understand this, notice that the ratio of first-best effort in the two cost states is given by $e_0^*(c_L)/e_0^*(c_H) = c_H/c_L$. Hence, as the ratio $c_H/c_L$ increases, there is a greater wedge between the levels of effort self-0 desires in each cost state. This wedge carries over to the levels of effort induced by the optimal goal: since self-0 desires relatively more effort in the low-cost state than in the high-cost state with an increase in the cost ratio, the optimal incomplete goal also induces relatively more effort in state $c_L$ than in state $c_H$. This is a testable prediction across tasks with variability in costs[5].

Finally, a marginal increase in $V$ also unambiguously leads to an increase in the level of goal deviation. The intuition for this is very similar to the reasoning for which an increase in the cost ratio increases $D_H$. To see this, computing the difference between the first-best levels of effort across the cost states yields $e_0^*(c_L) - e_0^*(c_H) = V(1/c_L - 1/c_H)$ which is increasing in $V$ since $c_L < c_H$. Again, the increase in the ex-ante optimal effort wedge carries over to the optimal incomplete goal and, thus, leads to a greater degree of goal deviation in the high-cost state.

[4]This substitutability between the two behavioral phenomena fits with the evolutionary argument of Suvorov and Van de Ven (2008) who argue that loss aversion may be an evolutionary response to time-inconsistent behavior, which allows individuals to more effectively self-regulate.

[5]Equivalently, one can perform the test with state-independent costs and a state-dependent benefit schedule, $\{V_j\}_{j=1}^n$.

## 3.5 Arbitrary, Finite State-Spaces

Now, consider the more general setup of the model with $\Theta = \{c_1, c_2, \ldots, c_n\}$ where $n$ is finite and $c_1 < c_2 < \ldots < c_n$. Assuming a general state-space complicates the analysis for a number of reasons. In particular, where in the binary state-space version of the model the individual could choose between only fully complete and fully incomplete goals, the domain of choice is now enriched to include goals that are only partially complete. However, as will be displayed in this section, incomplete goals still remain attractive self-regulatory devices in this general setting. This illustrates that the key results of simplified model presented in the previous section are robust to state-spaces with greater cardinality. Some results are also presented regarding when the individual is able to attain first-best utility in general and the optimal awareness set self-0 wishes to induce in order to achieve this.

### 3.5.1 Optimal Goals for a Highly Time-Inconsistent Individual

Suppose that the individual is very time-inconsistent with $\beta < \mu(0)$. As was the case in the binary state-space version of the model, the objective of self-0 in this region of the parameter space is to induce as large an effort profile as possible[6]. The following proposition displays the optimal goal for a highly time-inconsistent agent.

**Proposition 3.9.** *Suppose that $\beta < \mu(0)$ and $S(A) = A$ for all $A \in 2^{\Theta}$. Then, the optimal goal is fully incomplete and is given by*

$$g^* = (\{\Theta\}, \{c_1\}, \mu(1)\beta V/c_1)\,.$$

[6]Lemma A.1 stated in the Appendix shows that $\bar{e}_c \equiv \mu(1)\beta V/c$ is the maximum level of effort that can be implemented in any cost-state and $\bar{e}_c < V/c$ when $\beta < \mu(0)$.

*The individual falls short of her goal in cost-states* $c_2, c_3, \ldots, c_n$ *which occurs with ex-ante probability* $1 - P(c_1)$.

*Proof.* See Appendix. ■

This finding establishes that the finding of Proposition 3.5 holds in this more general setting. Specifically, the individual chooses a fully incomplete goal that is deviated from in all cost-states except $c_1$. Moreover, all goal deviation involves the individual falling short of this self-set expectation in states where deviation occurs. Hence, for a very time-inconsistent individual, the baseline result from the binary state-space model of the previous section continues to hold.

### 3.5.2 Optimal Goals for a Fairly Time-Consistent Individual

Now, consider the case in which $\beta \in (\mu(0), 1]$, such that the self-control problem of the individual is less severe than in the previous section. In this region of the parameter space, the problem is much more complicated due to the fact that overregulation now becomes a possibility: for a given goal, self-1 may now over-provide effort in some cost-states. In order to solve for the optimal goal, self-0 must balance the trade-offs inherent in choosing goals that lead to an under-provision of effort in some states and an over-provision in others. While this high-dimensional combinatorial problem does not offer a simple analytical solution, it is still feasible to derive some properties of the optimal goal at parameter values in which ex-ante optimal can be induced. Proposition 3.10 provides this characterization.

**Proposition 3.10.** *There exists a non-empty set of time-inconsistency parameters,* $\mathcal{B} \subset [\mu(0), 1]$*, for which self-0 is able to implement ex-ante optimal effort* ($e(c, g^*) =$

*$V/c$ for all $c \in \Theta$). The goal that achieves this, $g^* = (\xi^*, \phi^*)$, satisfies the following properties:*

*(a)* ***Minimal Awareness:*** *Self-1 is aware of at most two states: $|\xi^*| \leq 2$.*

*(b)* ***Extreme-Point Salience:*** *The optimal awareness set satisfies $\xi^* \subset \{c_1, c_n\}$.*

*Proof.* See Appendix. ■

This proposition provides interesting predictions regarding the states the individual optimally highlights. The *minimal awareness* property states that any goal in which ex-ante optimal effort is to be induced will leave self-1 aware of at most two events. Hence, large decreases in awareness are extremely useful to self-0 and, at least at points where ex-ante optimal effort is feasible, goal deviation should be *expected* rather than an *exception*.

The property of *extreme-point salience* states that the individual will only ever find it useful to make either the lowest cost-state or the highest cost-state, and all intermediate cost-states will only push her future self away from exerting ex-ante optimal effort. The intuition for this is as follows. The main issue self-0 faces is an under-provision of effort in low-cost states and an over-provision of effort in high-cost states. By grouping low cost-states (high-cost states) together and making the $c_1$ ($c_n$) salient, this trade-off is optimally balanced. At these knife-edge values of $\beta \in \mathcal{B}$, the trade-off is perfectly balanced and the individual is able to implement the ex-ante optimal effort profile[7].

[7]While the set $\mathcal{B}$ has Lebesgue measure zero, it can be shown that the described goal for $\beta \in \mathcal{B}$ will still be optimal for some open neighborhood around $\beta$.

## 3.6 Discussion

This section discusses some of the key assumptions and findings of the model and suggests some possible extensions. First, we extend the model to allow for the individual to abstain from any goal-setting. It is shown that this allows for more realistic predictions for fully time-consistent individuals (with $\beta = 1$), but does not affect the main predictions of the model. Second, we discuss the potential implications of placing constraints on the set of partitions, $\Pi$, that the individual is able to choose from. Finally, the implications of assuming that self-0 does not consider psychological utility when solving for the optimal goal are discussed.

### 3.6.1 Allowing for Goal-Abstention

Since the model promotes the sophistication of the decision-maker as a key advantage, it seems reasonable that goal abstention should be in self-0's domain of choice. For this extension, we focus on the simplified version of the model presented in Section 4 with a binary state space[8]. Define abstention from goal-setting, $g_\emptyset$, to be the *laissez-faire* policy that does not induce a reference point for self-1. Consequently, self-1 is free to implement her optimal level of effort provision if $g_\emptyset$ is utilized by self-0; that is, $e(c_j, g_\emptyset) = \beta V/c_j$ for $j = L, H$. Thus, abstaining from goal-setting does nothing to motivate the individual to overcome her self-control problem.

For very time-inconsistent agents, one would expect that abstention leads to ex-ante equilibrium utility far below first-best levels and some form of self-regulation would be beneficial. In contrast, as $\beta \to 1$, $e(c_j, g_\emptyset) \to e_0^*(c_j)$, $j = L, H$, and ex-ante optimal effort is provided without intervention from self-0. Since the previous analysis displayed that, except in a knife-edge case where $q = 1/2$, neither complete

[8]The results will also go through for a more general state-space, although the analysis is more complicated without adding any additional intuition.

nor incomplete goals can induce first-best effort provision as the self-control problem dissipates, it follows that one would expect goal abstention to be commonly observed behavior for relatively time-consistent individuals. Proposition 3.11 proves that this intuition is correct.

**Proposition 3.11.** *Suppose self-0 can now choose to abstain from goal-setting by $g_\emptyset$. Then, for all $q \in (0,1)$, $q \neq 1/2$, there exists a threshold, $\beta_\emptyset \in (\mu(0), 1)$, such that*

*(a) For $\beta < \beta_\emptyset$, it is optimal to use some form of goal-setting.*

*(b) For $\beta > \beta_\emptyset$, goal abstention constitutes optimal behavior.*

*Proof.* See Appendix. ■

First, notice that $\beta_\emptyset > \mu(0)$. Hence, the main predictions of the model persist even when allowing for goal abstention. The individual still finds it optimal to utilize incomplete goals for a non-trivial region of the parameter space. Indeed, if $\beta_\emptyset > \beta^*(q)$ for all $q$, then predictions regarding both the extensive and intensive margins of goal deviation from Section 4 are unaffected.

The main benefit of including goal abstention in the domain of choice for self-0 is realism: when restricted to use some form of goal-setting, the agent was, in general, unable to achieve first-best utility even as she approached full time-consistency. Proposition 3.11 suggests that, when $\beta$ is sufficiently large, self-0 will abstain from goal-setting and give full control over effort provision to self-1, which achieves first-best utility in the limit as $\beta \to 1$.

In summary, allowing for goal-abstention leads to the more realistic prediction that a fully time-consistent individual will undertake the ex-ante optimal levels of

effort provision. However, it does not eliminate the fact that incomplete goal-setting may constitute optimal behavior and consequently, the potential for goal deviation persists with this expansion of self-0's domain of choice.

### 3.6.2 No Psychological Disutility to Self-0

An assumption of the model is that self-0 does not incorporate the psychological utility that her future self will incur when choosing the optimal level of effort provision. This was assumed to ensure the focus of the analysis is on self-0 utilizing goals *instrumentally* (that is, to overcome the motivation problem that she faces) and abstract from any incentives at the planning stage to manage her future psychological utility. It also serves as a technical assumption that allows a much cleaner characterization of the optimal goal.

However, this assumption may be an oversimplification for a number of reasons. First, it is the same individual at both $t = 0$ and $t = 1$, it may be somewhat unreasonable that a sophisticated self-0 does not consider how she will feel when deviating from the optimal incomplete goal. Second, there may be a concern that, since she does not directly care about the loss aversion parameter, self-0 would desire this parameter to be as large as possible to optimally self-regulate. In regards to the first point, extending the model to allow for the possibility that self-0 directly incorporates psychological concerns into her utility may be a useful exercise. It is likely that this would decrease the extent to which incomplete goals constitute optimal behavior: since such goals are deviated from in equilibrium this will add an extra 'cost' of such goals into ex-ante welfare. However, it seems reasonable to think that the propensity to set incomplete goals will not disappear - there will just be a tradeoff. Moreover, if the model were to be applied to an *interpersonal* environment (such as a contracting setting), then it would not be desirable to incorporate psychological concerns into a

self-interested principal's utility.

With respect to the second concern, there is in fact no problem. Indeed, recall that $\lambda$ shares similar comparative statics to the time-inconsistency parameter, $\beta$. In particular, for $\beta$ in a neighborhood of $\mu(0)$, ex-ante welfare is initially increasing in $\lambda$, before beginning to decrease in $\lambda$. Hence, there is a sense in which self-0 *indirectly* desires a finite $\lambda$ already because the degree of loss aversion affects the effort provision decision of self-1. To illustrate, consider the model of Section 4 with a binary state-space. Then, in the case of an incomplete goal, when self-1 is already over-providing effort in the high cost state, an increase in $\lambda$ will only exacerbate this sub-optimality. Indeed, fixing $\beta$, the $\lambda^*$ that solves

$$\beta = \mu(0; \lambda^*) \equiv \frac{1+\eta}{1+\eta\lambda^*} = \beta$$

achieves ex-ante optimal utility.

## 3.7 Conclusion

A novel micro-foundation for goal-setting, with goal deviation in equilibrium, has been provided. The key concept that drives the results is that the individual has endogenous control over her awareness: the type of goal she uses, and visualizes, influences her future perceptions of the state-space, by affecting the set of states she will be aware could have possibly occurred. This captures the idea that planning and doing are different mental conditions. When planning, an individual can rationally think through all possible contingencies and plan accordingly. Alternatively, when doing, they are focused on completing the task in the realized state and are only able to recall counterfactual states that were explicitly planned for at the goal-setting

stage.

This formulation of endogenous awareness allows for different types of goals to be distinguished. In particular, goals that induce large degrees of unawareness (incomplete goals) generate distinct behavioral responses from goals that induce full awareness (complete goals). This is because full awareness leads the individual to compute a stochastic reference-point that includes costly *cross*-state utility comparisons. Incomplete goals, in contrast, reduce psychological costs, relax incentive constraints, and hence, lead to greater levels of implementable effort. In the binary state-space version of the model, such goals were shown to be optimal except for the case in which the individual is sufficiently time-consistent such that a complete goal does eventually constitute an optimum. This is because the only feasible optimal incomplete goal is a blunt instrument for self-regulation that induces an over-provision of effort in the high-cost state, which is magnified as the individual's self-control problem dissipates. The incentives to utilize incomplete goals were also shown to be robust to the introduction of a more general state-space. Moreover, the general model provided additional predictions regarding how moderately time-inconsistent individuals optimally compartmentalize the state-space. In particular, it was shown that, for a subset of time-inconsistency parameters, the individual could achieve first-best utility by grouping cost-states into at most two distinct groups, and making extreme points of the cost-distribution salient.

The model also provides novel comparative statics with respect to each of the behavioral parameters. It is shown that there is a non-monotonic relationship between both goal deviation and the individual's willingness-to-pay for full commitment and the degree of time-inconsistency. In particular, individuals with intermediate levels of self-control find incomplete goals to be extremely effective self-regulatory

tools. This may help to explain why the take-up of commitment devices is small in practice: soft commitment in the form of reference dependence and loss aversion may already be successful deterrents of giving in to temptation. Moreover, the it is shown that, at the point at which goal deviation is maximized, willingness-to-pay is equal to zero. This suggests that paternalistic policy prescriptions that focus on helping individuals to unconditionally achieve their self-set goals may be somewhat misguided. Empirically testable predictions are also generated with respect to loss aversion, the variation in costs, and the task's benefit. For each of these parameters, there are well-known experimental methods for their elicitation. Therefore, the model should be more easily validated or rejected than some of the alternative explanations for goal deviation offered in the literature.

This model of endogenous awareness may prove insightful for a wide range of economic problems. First, it may be useful to apply this theory in a more standard contracting environment. The literature has already investigated the implications of unawareness in a range of contracting environments (Von Thadden and Zhao (2012), Filiz-Ozbay (2012), Tirole (2009)). In Von Thadden and Zhao (2012) the agent is unaware of the possible actions that he or she may take, and in Filiz-Ozbay (2012), the individual is aware of only a subset of the states of the world. On the other hand, Tirole (2009) argues that it is cognitively costly to be think through all contigencies, design covenants, and to be aware of all of the implications of a contract. All of these papers find that their specific forms of unawareness are a micro-foundation for incomplete contracting in their respective markets. The present paper shows that endogenous awareness can also affect the incentives for effort provision through stochastic reference-point calculations. Since stochastic reference-point calculations with full awareness depress effort, a principal may wish to maintain an agent's unawareness through incomplete contracts (which could be modeled in a similar

fashion to the incomplete goals in this paper). Therefore, this theory may provide an alternative explanation for incomplete contracts and their structure in markets. Moreover, the formulation of the model was able to capture deviation and thus, may prove insightful for understanding the empirical observation that contract terms are sometimes not adhered to.

Second, the theory should yield new, valuable insights for organizational economics. Most organizations set goals and, like individuals, they often fall short of these internally set "expectations". Moreover, when states of the world realize in which organizational goals are deviated from, firms are often punished with depressed share prices and managerial turnover, which suggests that markets attribute deviation from targets to some form of incompetence. However, applying our theory of endogenous awareness, a more sophisticated explanation for this phenomena may be that there are adverse consequences when organizations plan for every contingency: if workers (or shareholders) exhibit stochastic reference-dependence, then it may be the case that such planning is sub-optimal as it tightens incentive constraints which may adversely affect firm productivity (or its share price). Thus, the observation of non-state-contingent planning by organizations may be rationalized by this theory. These potential directions will be explored in future research.

# Chapter 4

# Communication with Endogenously Naive Receivers

## 4.1 Introduction

The main motivation for the theoretical model provided in this chapter is as follows: (a) whether individuals are naive or sophisticated affects their ability to discover the informational content of observed messages, (b) transmitters of information may try to take advantage of such naiveté, and (c) the likelihood with which an individual is naive or sophisticated should be a function of the context in which information is generated.

In order to achieve this, we will focus on a dynamic model of cheap-talk communication as in Crawford and Sobel. A strategic sender observes the state of the world and chooses a message conditional on this state. The message profile (as a function of all states) implies some true joint distribution between messages and states. A sophisticated receiver recognizes that the sender is strategic and, thus, uses this true joint distribution to determine the optimal action. In contrast, if the receiver

is naive then she will hold an incorrect perception of the correlation between messages and the state. In particular, a naive receiver mistakenly believes that there is some probability with which the sender is honest and transmits information truthfully.

Whether the receiver ends up sophisticated or naive will be a function of the mode of cognition she ends up in. Specifically, the receiver being in a high cognitive-state is equivalent to being sophisticated. On the other hand, if the receiver is in a low cognitive-state, then this is equivalent to her being naive and, therefore, she will hold incorrect beliefs. Cognition is modeled as an active choice that the individual makes which is determined according to the following trade-off: being in a state of high cognition allows the receiver to make a more informed decision but such sophistication comes at a cognitive cost.

Given the setup of the model, it is established that the rational benchmark (where the receiver correctly perceives that the sender is strategic) has a unique equilibrium in which no information is transmitted. It is shown, however, that the unique equilibrium with bounded cognition can involve a strictly positive amount of information transmission. Hence, bounded cognition can play a role in ensuring the truthful revelation of information. The reasoning for this is that a strategic sender will have a strong incentive to tell the truth to a receiver with bounded cognition due to the fact that such a receiver will then think that she is honest with a high probability in future periods and, as such, will see no reason to invest in cognition to discover the sender's true type.

An implicit characterization of the extent to which the sender tells the truth is provided. Using this implicit characterization, we are able to establish a *paradox of cognition*: the higher the cognitive ability of the individual (the lower are cognitive

costs), the less information is transmitted in equilibrium. This implies that a receiver may actually benefit from being perceived by the sender as being cognitively disadvantaged: a higher degree of truth-telling will allow her to correlate her actions more effectively with the state.

This chapter proceeds as follows. In the following section, a small review of the key related literature is provided. In section 3, the model is detailed. Section 4 provides the main results, while section 5 concludes. All proofs are relegated to the appendix.

## 4.2 Related Literature

There is an expansive literature on communication in economics (see Sobel (2013) for an extensive review). Within this literature, the work in this chapter is most closely related to models that build on the pioneering work of Crawford and Sobel (1982) by allowing the sender to have different types. Some key examples of such work include Benabou and Laroque (1992) and Morris (2001). In contrast to these papers, the work here also allows for there to be heterogeneity in terms of the realized type of the receiver, depending on whether she ends up naive or sophisticated. It is shown that this permits equilibria that would not be present if all uncertainty was determined by the sender's type.

The work is also closely related to models of communication in which a fraction of the population is naive with respect to the information that is being transmitted (Kartik, Ottaviani, and Squintani (2007), Ottaviani and Squintani (2006), Chen (2011)). These papers show that information transmission can be permitted by the presence of naive receivers, and this is achieved by biasing messages upward to take

advantage of such receivers. Similarly here, bounded cognition is shown to be a force that permits the truthful revelation of information. In addition, here the fraction of the receiver population that is naive will be responsive to the incentives within the model. Hence, the model also provides predictions on how the receiver's naiveté evolves over time. This adjustment in the proportion of the population that is naive also proves necessary for information revelation: if this proportion were fixed as in the prior literature, then the model presented here would predict that the sender always lied in equilibrium in every time period.

## 4.3 The Model

There are two time periods, denoted $t = 1, 2$. A long-run sender ($S$) faces a sequence of short-lived receivers ($R$), that live for only one period. In each $t$, a state-of-the-world, $\omega_t$, is drawn from the state-space $\{0, 1\}$. Draws of $\omega_t$ are independent and identically distributed across time. Let $\pi$ denote the probability that $\omega_t = 1$.

**Strategies:** The model of communication here is similar to the cheap talk setting of Crawford and Sobel. The sender observes $\omega_t$ in each time period and must decide if and how to communicate this information to the receiver. Specifically, a sender of type $\omega_t$ must choose a message $m \in \{0, 1\}$ to transmit to the receiver. Let $m_t = (m_t(0), m_t(1))$ denote the strategy profile for whole types of the sender. Then $m_t$ summarizes whether there is any information contained in a specific message. Given $m_t$, the period-$t$ receiver chooses some action $a \in [0, 1]$. It is assumed that, at the end of $t = 1$, the receiver in $t = 2$ observes the state $\omega_t$.

**Preferences:** The chosen action affects the payoffs of both the sender and the receiver. We assume that the payoff to the sender in time $t$ is independent of $\omega_t$

and is given by $u_S(a) = a$. Hence, the sender wants the receiver to take as high an action as possible. The long-lived sender values a stream of payoffs using a discount rate of one. In contrast, the objective of the receiver is to match her action to the state, $\omega_t$. In particular, receiver preferences are given by $u_R(a, \omega_t) = -(a-\omega_t)^2$.

**Endogenous Naiveté:** There are two 'types' of receiver: naive or sophisticated. A sophisticated receiver is able to deduce that the sender is strategic and, hence, chooses an action conditional on the informational content of the message received. A naive receiver, on the other hand, mistakenly believes that there is some chance that the sender is *honest* and transmits information truthfully according to the rule $m_H(\omega_t) = \omega_t$. A key idea here is that the likelihood the receiver ends up either naive or sophisticated will be a function of how much she decides to invest in cognition.

Formally, let $\mu_t$ denote the probability that the period-$t$ receiver believes that the sender is strategic (where $1 - \mu_t$ denotes the mistaken probability that the sender is honest). Then, conditional on observing message $m$, the receiver chooses a cognitive strategy $\rho \in [0, 1]$, where $\rho$ denotes the probability that the receiver learns the sender's true type and $1 - \rho$ denotes the probability the receiver learns nothing. It is assumed that investment in cognition is costly, with cost-function $T(\rho) = \kappa\rho^2/2$. This cost function satisfies many intuitive properties of cognition: (a) cognition is costly ($T(\rho) > 0$); (b) higher levels of cognition are increasingly costly ($T'(\rho) > 0$); and the marginal cost of cognition is increasing ($T''(\rho) > 0$). The gains to cognition will be endogenously determined in equilibrium and will come from a reduction in uncertainty from learning the sender's true type.

With this formulation, we can interpret $\rho$ to be likelihood that the individual ends up being sophisticated (with corresponding value $1 - \rho$ representing the probability

the receiver ends up naive). Taking a population interpretation (where there are many identical receivers), we can interpret $\rho$ to be the *fraction of the population* that is sophisticated. We are interested in how $\rho$ evolves over time.

**Receiver's Optimization Problem:** Let $i \in \{S, H\}$ denote the set of types that receiver perceives the sender could be. If the receiver is sophisticated, then she observes that the sender is strategic. Hence, conditional on observing message $m$, she chooses her action to solve

$$a_S(m) \in \arg\max_a E[u_R(a, \omega_t)|m, i = S, m_t]$$

where $i = S$ and $m_t$ summarize the true informational content of message m. Given that the utility function is quadratic, it is simple to show that the solution to this problem is given by $a_S(m) = Pr(\omega_t = 1|m, i = S, m_t)$.

In contrast, if the receiver is naive, then she does not observe the type of the sender after her cognitive sampling procedure. Hence, the naive receiver chooses her action, conditional on message $m$, to solve

$$a_N(m) \in \arg\max_a E[u_R(a, \omega_t)|m, m_t]$$

where the only information that the naive receiver can condition on is perceived amount of information being revealed by the strategic sender in equilibrium, $m_t$. Again, it is simple to show that the solution to this problem yields $a_N(m) = Pr(\omega_t = 1|m, m_t)$.

Given the optimal actions of both the naive and the sophisticated receiver, it is now possible to formulate the cognition-optimization problem of the receiver. Specifically,

conditional on message $m$, she chooses to invest in the level of cognition, $\rho^*(m)$, that solves

$$\max_{\rho} \quad \rho\mu_t E[u_R(a_S(m), \omega_t)|m, i = S, m_t] + (1-\rho)E[u_R(a_N(m), \omega_t)|m, m_t] - \kappa\frac{\rho^2}{2}.$$

**Equilibrium:** We proceed via backwards induction, using the solution concept of Bayesian Nash equilibrium to solve the game in each period. Some extra assumptions are also placed on the set of equilibria considered. First, we will only consider equilibria in which the sender chooses to be truthful when observing that the state is equal to one; that is, $m_t(1) = 1$. This restriction allows us to focus on the extent of truth-telling by the strategic sender when $\omega_t = 0$ is observed as an important variable of interest. It also ensures that the receiver always knows that the state is $\omega_t = 0$ if $m_t = 0$ is observed and, thus, will have no need to invest in cognition upon observing this low message.

The second assumption is related to the information set of the receiver in $t = 2$. Since the receivers are short-lived, then it will be assumed that the receiver in $t = 2$ does not have access to any of the information that the $t = 1$ receiver may have learned by investing in cognition. In particular, the $t = 2$ receiver will not automatically become sophisticated if the $t = 1$ receiver is sophisticated, but rather will have to undergo her own cognitive investment. Therefore, the only available information to the $t = 2$ receiver is whether or not the sender has lied in $t = 1$ (i.e. she observes $m_1$ and $\omega_1$), which she will use to update her (mistaken) belief that the sender type is strategic.

## 4.4 Main Results

In this section, a characterization of the equilibrium under different parameterizations of the model is provided. In order to do this, we proceed through the following steps. First, we derive the cognitive best-response function of the receiver in a given period $t$, conditional on observing $m = 1$, for an arbitrary probability the sender lies and belief that the sender is strategic. Using this, we then characterize the equilibrium in $t = 2$ and show that the sender that observes $\omega_2 = 0$ will always lie in this period. Finally, we use the equilibrium in $t = 2$ to characterize the incentives for the sender to tell the truth in $t = 1$ and provide relevant comparative statics of this equilibrium.

### 4.4.1 Cognitive Best-Response

Given that we are only considering equilibria with $m_t(1) = 1$, it follows that the receiver knows that the state is zero conditional upon observing a message of zero. Given this, $a_S(0) = a_N(0) = 0$, and the receiver earns the same utility regardless of whether naive or sophisticated. Since cognition is costly, it follows that $\rho_t(m = 0) = 0$ and the agent will always be in a naive state conditional on observing the low message.

It is more interesting to focus attention on how cognition responds to incentives conditional on observing $m = 1$. The following lemma provides an explicit computation of this best-response function.

**Lemma 4.1.** *Let $(\mu, \nu)$ be a vector where $\mu$ is receiver's belief that the sender is strategic, and $\nu$ is the belief that the strategic sender tells the truth. Define $f(\mu, \nu) \equiv Pr(\omega_t = 1 | m = 1, m_t)$ and define*

$$\bar{\kappa}(\mu, \nu) = \mu(1 - \nu)\frac{1 - \pi}{\pi}\left[f(\mu, \nu)^2 - f(1, \nu)^2\right].$$

*Then, the cognitive best-response of the receiver conditional on observing $m = 1$ is*

$$\rho(\mu, \nu) = \min\left\{\frac{\bar{\kappa}(\mu, \nu)}{\kappa}, 1\right\}.$$

*$\rho(\mu, \nu)$ is inverse U-shaped in $\mu$ and decreasing in $\nu$.*

*Proof.* See Appendix. ■

The above lemma provides an explicit computation of the cognitive best-response of the receiver conditional on observing the high message. Looking at the expression for $\rho(\mu, v)$, we see that, when $\kappa$ is sufficiently small, the individual will choose to maximally invest in cognition (i.e. set $\rho = 1$), and cognition is decreasing in $\kappa$ for a given $(\mu, \nu)$.

Lemma 4.1 establishes two important comparative statics on the optimal choice of cognition for the receiver. The first is that cognition is decreasing in the probability that the sender tells the truth ($\nu$). The reason for this is that the more likely it is that the sender is telling the truth, the more likely the true state is actually $\omega_t = 1$ conditional on observing a message of one. Hence, the value of cognition is decreasing the more likely the sender tells the truth.

The second is that the cognitive best-response is inverse U-shaped in the receiver's belief that the sender is strategic, $\mu$. This is intuitive: the more extreme $\mu$ is, the more confident the receiver is that the sender is of a particular type. Given that a naive receiver is already tailoring her action to her belief regarding the sender, this implies that the value of discovering whether the sender is strategic or not is also lower. Thus, cognition is decreasing the more extreme $\mu$ is.

It is useful to use Lemma 4.1 to establish how the per-period profit of the sender, conditional on the receiver observing $m = 1$, is affected by $(\mu, \nu)$. Understanding these effects will be useful for characterizing the equilibrium in the following section.

**Lemma 4.2.** *The period $t$ profits of the sender for an arbitrary $(\mu, \nu)$, conditional on the receiver observing $m = 1$, are given by*

$$\varphi(\mu, \nu) = \rho(\mu, \nu) f(1, \nu) + (1 - \rho(\mu, \nu)) f(\mu, \nu).$$

*This expression is increasing in $\nu$ and decreasing in $\mu$ and is bounded below by $\pi$.*

*Proof.* See Appendix. ■

The profits of the sender in period $t$ are simply the expected value of the receiver's action, where this expectation is taken with respect to the likelihood the receiver is naive or sophisticated. When the receiver is sophisticated, she observes that the sender is sophisticated and, based on an earlier argument, chooses $Pr(\omega_t = 1 | m, i = S, m_t) = f(1, \nu)$. In contrast, when the receiver is naive, she does not observe whether the sender is strategic or not and, as such, computes the conditional probability that the state is equal to one using only the information contained in the message: $Pr(\omega_t = 1 | m, m_t) = f(\mu, \nu)$.

Lemma 4.2 establishes that $\varphi$ is increasing in the probability that the sender tells the truth. The intuition for this is as follows. As the sender tells the truth more often then, conditional on the receiver observing $m = 1$, she will both choose a higher action (she thinks it is more likely that the state is $\omega = 1$) and choose to invest less in cognition (by Lemma 4.1) which increases profits to the sender.

The comparative static generated with respect to $\mu$ is, however, slightly more nuanced. Lemma 4.2 shows that, as $\mu$ increases, $\varphi$ decreases. In order to understand this, it is important to observe that there are two (potentially competing) effects. First, as $\mu$ increases the receiver becomes more certain that the sender is strategic and decreases her action when naive ($f(\mu, \nu)$ decreases). Second, as $\mu$ increases, the cognitive strategy of the receiver can either increase or decrease (Lemma 4.1). In the region that $\rho(\mu, \nu)$ increases, then these two effects move in the same direction, leading to a decrease in profits. In contrast, when $\rho(\mu, \nu)$ the two effects are in opposite directions. It can be shown, however, that the first effect dominates the second and profits to the principal decrease overall.

These two comparative statics on $\varphi$ will be important for proving that the equilibrium has a particular structure, as will be shown in the following sections.

### 4.4.2 Period $t = 2$ Equilibrium

Suppose that we have entered period $t = 2$ and that the receiver believes that the sender is strategic with probability $\mu_2$. As is relatively standard in communication games of this form, since this is the final period, when $\omega_2 = 0$, the sender will always have a strict incentive to lie and, as such, the unique equilibrium will have $m_2(0) = 1$. In particular, since the receiver is always certain that $\omega_2 = 0$ when $m = 0$ based on our equilibrium restrictions, the sender will receive a payoff of zero from choosing message $m = 0$. Since $\varphi(\mu, \nu) > 0$ for any $(\mu, \nu)$ (one can easily show that $\varphi(\mu, \nu)$ is bounded below by $\pi > 0$), it follows that the $m(0) = 1$ must be the unique equilibrium in $t = 2$. This result is summarized in the Lemma 4.3.

**Lemma 4.3.** *Let $\mu_2$ denote the belief of the receiver that the sender is strategic. Then, the unique $t = 2$-equilibrium has $m_2(0) = 1$ with probability one. The cognitive*

*strategy of the agent is given by* $\rho(\mu_2, 0)$ *and the profits of the sender, irrespective of the true value of* $\omega_2$*, are given by* $\varphi(\mu_2, 0)$*.*

A proof is not required since the argument preceding the statement of Lemma 4.3 is sufficient. The remaining components of the characterization come from noticing that if $m_2(0) = 1$ with certainty, then $\nu = 0$ and, as such, the equilibrium cognitive strategy and sender profits can be determined from the expressions derived in Lemma 4.1 and Lemma 4.2 with $(\mu, \nu) = (\mu_2, 0)$.

This result establishes that a dynamic setting (with multiple time periods) is necessary if there is to be any truthful revelation of information, as the one shot equilibrium leads to the sender always choosing a message of 1. From the perspective of the sender in $t = 1$, $\varphi(\mu_2, 0)$ will be the continuation value of taking different strategies. Since Lemma 4.2 established that this value is decreasing in $\mu_2$, this implies that the sender may have strong incentives to *tell the truth* if the receiver is expecting her to lie. More strongly, while 'always lie' will be established to be the equilibrium in the rational benchmark, a condition will be provided such that this is *never* an equilibrium when the cognitive cost parameter of the agent is strictly greater than zero.

### 4.4.3 Period $t = 1$ Equilibrium

This section contains the results of most interest for this chapter. The main variable of interest is the equilibrium probability with which the sender who observes $\omega_1 = 0$ tells the truth (i.e. sends the message $m_1(0) = 0$). We first establish what the equilibrium in the rational benchmark is. Here, the rational benchmark is a receiver with $\kappa = 0$, such that she can always observe that the sender is strategic for free. Lemma 4.4 shows that the unique equilibrium in the rational benchmark involves

the sender always lying conditional on observing $\omega_1 = 0$.

**Lemma 4.4.** *Suppose that $\kappa = 0$ so that the receiver always observes that the sender is strategic. Then, the unique equilibrium is $m_1(0) = 1$ with probability one.*

*Proof.* See Appendix. ■

Combining Lemma 4.4 and Lemma 4.3, one gets the prediction that no information can be transmitted when the receiver is rational and, therefore, able to observe that the sender is strategic. Hence, in this framework if one observes information transmission, then it can be attributed to the fact that the receiver finds investment in cognition costly.

Before proceeding to find conditions such that information transmission can occur in equilibrium, it is useful to establish when this is not the case. Lemma 4.5 shows that, when the ex-ante probability that the $\omega_t = 1$ is sufficiently high, then the unique equilibrium is the same as that in the rational benchmark.

**Lemma 4.5.** *Suppose that $\pi \geq 1/2$. Then, the unique equilibrium is $m_1(0) = 1$ with probability one and no information is transmitted in equilibrium.*

*Proof.* See Appendix. ■

Lemma 4.5 shows that no information can be transmitted in equilibrium when $\pi$ is sufficiently large, regardless of the cognitive cost of the receiver. The intuition for this result is simple: when the sender lies in $t = 1$, then his period $t = 1$ profits are bounded below by $\pi$ (Lemma 4.2) and his period $t = 2$ profits are equal to $\pi$ (the receiver in $t = 2$ observes the lie and there knows the sender is strategic). In

contrast, if the sender deviates and tells the truth, then given the specification of this equilibrium, the receiver in $t = 2$ will believe that the sender is honest with probability one (since she was expecting the strategic sender to lie and choose $m = 1$ with probability one). This gives a current period payoff of zero and a maximal payoff of one in $t = 2$. When $\pi \geq 1/2$, this deviation is not profitable and the unique equilibrium is the same as that in the rational benchmark.

We now want to derive conditions such that equilibrium involves some transmission of information. As a first step, it is useful to check whether always telling the truth can ever constitute an equilibrium. Propostion 4.1 establishes that this can never be the case: irrespective of the parameters of the model, the sender will never be willing to choose $m_1(0) = 0$ with probability one in equilibrium.

**Proposition 4.1.** *There is no equilibrium with $m_1(0) = 0$ with probability one.*

*Proof.* See Appendix. ■

Given this, one should never expect to observe a perfect transmission of information. The intuition for this is as follows: if the receiver in $t = 1$ is expecting the sender to tell the truth with probability one, then conditional on observing $m = 1$, she will choose not to invest in cognition and choose the highest action of one. This incentive to deviate proves to be greater than the incentive to maintain the receiver's uncertainty as to whether he is a strategic sender or not and, therefore, full truth-telling is not sustainable in equilibrium.

Given this, if there is to be any transmission of information in equilibrium, it must only be partial: that is the sender must mix between telling the truth ($m_1(0) = 0$) and lying ($m_1(0) = 1$). Lemma 4.5 has already established that no

transmission of information can occur when $\pi \geq 1/2$. Hence, for the remainder we assume that $\pi < 1/2$. In addition, we assume that $\mu_1 > \pi/(1-\pi)$, such that the receiver's (incorrect) belief that the sender is strategic is sufficiently large. Under this condition, one can show that the unique equilibrium is mixed and a characterization of this is provided in Proposition 4.2.

**Proposition 4.2.** *Fix $\kappa > 0$, $\pi < 1/2$ and $\mu_1 > \pi/(1-\pi)$. Then, in the unique equilibrium the sender mixes between choosing $m_1(0) = 0$ and $m_1(0) = 1$ with probability $\nu^* \in (0,1)$. This equilibrium mixing probability is characterized by the implicit function*

$$\varphi(\mu_1, \nu^*) + \pi = \varphi(\mu_2, 0)$$

*where*

$$\mu_2 = \frac{\mu_1 \nu^*}{\mu_1 \nu^* + (1 - \mu_1)}$$

*is the $t = 2$-receiver's Bayesian update on the sender's type upon observing $m_1 = 0$.*

*Proof.* See Appendix. ■

This proposition provides a sufficient condition that ensures that there is a positive amount of information transmission in equilibrium for *any* value of the receiver's marginal cost of cognition. This stands in contrast to the rational benchmark in which no transmission could occur when the strategic nature of the sender was observed by the receiver. Therefore, the informativeness of the equilibrium is a function purely of the bounded cognition of the receiver.

The intuition for the proposition is as follows. First, it is simple to show that, under the conditions specified, there can be no equilibrium in which the sender always lies, due to the fact that the incentives to deviate and tell the truth are large.

These incentives are driven by the fact that the receiver in $t = 2$ will believe with probability one that the sender is non-strategic upon observing that a message of zero was sent in $t = 1$, which leads such a receiver to be naive with probability one and choose the highest possible action ($a_2 = 1$). Combining this with Proposition 4.1, this implies that there must be a mixed equilibrium. In such an equilibrium, the $t = 1$-sender that has observed $\omega_1 = 0$ must be indifferent between sending $m = 0$ and $m = 1$, which is precisely what the condition $\varphi(\mu_1, \nu^*) + \pi = \varphi(\mu_2, 0)$ ensures.

Hence, in this communication game we have established that bounded cognition is sufficient (under some conditions) for ensuring that information can be transmitted in equilibrium. A remaining question of interest is the extent to which truth-telling is affected by the cognitive ability of the agent (i.e. the marginal cost of cognition parameter, $\kappa$). The following proposition establishes that the extent of truth-telling ($\nu^*$) is actually *increasing* in the marginal cost of cognition, $\kappa$.

**Proposition 4.3.** *Let $\nu^*(\kappa)$ denote the equilibrium probability that the sender tells the truth as derived from Proposition 4.2. Then, $\nu^*(\kappa)$ is strictly increasing in $\kappa$.*

*Proof.* See Appendix. ■

Proposition 4.3 displays a *paradox of cognition:* as the receiver's cognitive ability *decreases*, the sender will actually choose to transmit *more* information in equilibrium. In the extreme, as $\kappa \to 0$, no information is transmitted in equilibrium whatsoever (Lemma 4.4).

The reasoning behind this somewhat paradoxical finding is as follows. In order for the equilibrium condition described in Proposition 4.2 to be satisfied, it must be the case that $\rho(\mu_2, 0) < 1$ (which is equivalent to there being some benefit to the sender

of choosing to reveal the state truthfully). The lower is $\kappa$, the greater the incentive of the receiver to invest in cognition. As a result, a high-cognitive ability sender needs to be relatively sure that the sender is non-strategic in order to not be willing to maximally invest in cognition in $t = 2$. This can be achieved only when $\nu^*$ is small such that it was very unlikely that a truthful message came from a strategic sender. As $\kappa$ increases, the receiver will decrease investment in cognition, which makes it more profitable to deviate in $t = 1$. In order to ensure that deviation does not occur, it follows that $\nu^*$ needs to increase (recall that $\varphi(\mu, \nu)$ was increasing in $\nu$). Thus, the equilibrium extent of truth-telling is increasing in the marginal cost of cognition.

## 4.5 Conclusion

In this short chapter, we have embedded the theory of endogenous modes of cognition provided in Chapter 2 in a communication setting. This has allowed us to model the naiveté or sophistication of an individual in a communication setting as an endogenous process that responds to the information strategy of the sender. Therefore, it provides a framework for thinking about the types of settings involving communication where we might expect to observe more or less strategic sophistication.

The main result of this chapter has been to show that, under specific parameterizations of the model, the unique equilibrium involves some information transmission, even though the rational benchmark would not allow for any such revelations. This implies that bounded cognition can be sufficient for ensuring that that the state of nature be revealed with positive probability. Interestingly, the model also predicts a paradox of cognition: the higher the cognitive ability of the receiver, the lower is the extent of truth-telling in equilibrium.

Of course, one of the driving forces of the predictions in this setting is that the receiver can only use cognition to determine whether or not the sender is strategic, and not to also discover additional information regarding the state. In future work, it would be interesting to think about how a receiver would trade-off gathering information along these two dimensions. It would also be useful to investigate an infinite horizon version of this model to establish whether truth-telling can be sustained in the long-run. Such avenues are left for future investigation.

# Appendix A

# Proof of Results

## A.1 Proof of Results in Chapter 2

### A.1.1 Proof of Lemma 2.1

Fix the cognitive strategy of the agent, $\rho$ and a frame $f$ which induces beliefs $f_P$. The principal has three choices (a) choose a contract that only $R$ buys, (b) choose a contract that only $F$ buys, and (c) choose a contract that both types buy. The most that can be made out of an $R$ type solves the following problem:

$$\max_{q,p} p - c\sum_{\omega} P(\omega)\frac{q(\omega))^2}{2}$$

subject to $\sum_{\omega} P(\omega)q(\omega) - p \geq 0$. Obviously, the individual rationality constraint binds and the resulting first order condition is given by

$$[q(\omega)] : P(\omega)\theta = P(\omega)cq_R(\omega)$$

or $q_R(\omega) = \theta/c$ for all $\omega \in \Omega$. Using the participation constraint, one gets $p_R = \theta^2/c$. Moreover, the framed type will also purchase this contract (since it offers constant consumption across states). Hence, both types purchase this contract and it yields profits to the principal

$$\pi(C_R) = \frac{\theta^2}{c} - \frac{c}{2}\sum_{\omega} P(\omega)\left(\frac{\theta}{c}\right)^2 = \frac{\theta^2}{2c}.$$

This obviously dominates any contract in condition (a).

The only thing required is to consider condition (b). Consider the relaxed problem

$$\max_{q,p} p - c\sum_{\omega} P(\omega)\frac{q(\omega)^2}{2}$$

subject to $\sum_{\omega} f_P(\omega)q(\omega) - p \geq 0$. Again, the individual rationality constraint obviously binds and the first order conditions are

$$[q(\omega)] : f_P(\omega)\theta = P(\omega)cq_F(\omega)$$

which yields $q_F(\omega) = \frac{\theta f_P(\omega)}{cP(\omega)}$ for all $\omega \in \Omega$. Using the participation constraint of the framed type, the price can be computed to be

$$p_F = \frac{\theta^2}{c} \sum_{\omega \in \Omega} \frac{f_P(\omega)^2}{P(\omega)}$$

The rational type will not be willing to purchase this contract, except when $f_P = P$. To see this, compute $U_R(C_F, f)$ to be

$$U_R(C_F, f) = \frac{\theta^2}{c} \sum_{\omega \in \Omega} P(\omega) \frac{f_P(\omega)}{P(\omega)} - \frac{\theta^2}{c} \sum_{\omega \in \Omega} \frac{f_P(\omega)^2}{P(\omega)} = \frac{\theta^2}{c} \left[ 1 - \sum_{\omega \in \Omega} P(\omega) \left( \frac{f_P(\omega)}{P(\omega)} \right)^2 \right] \tag{A.1}$$

Since $f(x) = x^2$ is strictly convex, it follows that, for $f_P \neq P$,

$$\sum_{\omega \in \Omega} P(\omega) \left( \frac{f_P(\omega)}{P(\omega)} \right)^2 > \left( \sum_{\omega \in \Omega} P(\omega) \frac{f_P(\omega)}{P(\omega)} \right)^2 = 1.$$

Hence, for $f_P \neq P$, $U_R(C_F, f) < 0$ and the rational type will not be willing to purchase $C_F$.

It follows that $C_R$ and $C_F$ are the only contracts offered in any equilibrium, where $C_R$ is purchased by both types if offered and $C_F$ is purchased only by the framed-type of the agent. ■

### A.1.2 Proof of Lemma 2.2

If the principal offers the contract $C_R(f)$, then both types of the agent will participate (Lemma 2.1). The profit to the principal on this contract is given by

$$\pi(C_R(f), f) = \frac{\theta^2}{c} - c\sum_{\omega\in\Omega} P(\omega)\frac{(\theta/c)^2}{2} = \frac{\theta^2}{2c}.$$

If instead, the principal offers the contract $C_F(f)$, then only the framed-type buys and profits are given by

$$\pi(C_F(f), f) = (1-\rho(f))\frac{\theta^2}{2c}\sum_{\omega\in\Omega}\frac{f_P(\omega)^2}{P(\omega)} = (1-\rho(f))\frac{\theta^2}{2c}\left(D(f_P||P)+1\right).$$

It follows that $\pi(C_R(f), f) \geq \pi(C_F(f), f)$ if and only if

$$1 \geq (1-\rho(f))(1+D(f_P||P)) \Leftrightarrow \rho(f) \geq \frac{D(f_P||P)}{1+D(f_P||P)}.$$

As argued in the text of the paper preceding Lemma 2.2, $\rho(f)$ can not be strictly greater than this threshold. We first search for an equilibrium with $\rho(f) < D(f_P||P)/(1+D(f_P||P)$. In such an equilibrium, $C_F(f)$ is offered with probability one and the optimal cognitive strategy solves

$$\max_{\rho\in[0,1]} (1-\rho)U_R(C_F, f) - T(\rho)$$

where $U_R(C_F, f)$ is given in equation (A.1). This gives the unconstrained first-order condition

$$\rho(f) = \frac{\theta^2}{\kappa c}D(f_P||P). \tag{A.2}$$

Note that $\rho(f)$ satisfies the necessary inequality if and only if

$$\frac{\theta^2}{\kappa c}D(f_P||P) < \frac{D(f_P||P)}{1+D(f_P||P)} \Leftrightarrow \kappa > \frac{\theta^2}{c}(1+D(f_P||P)).$$

Define $\bar{\kappa}(f) \equiv \frac{\theta^2}{c}(1 + D(f_P||P))$. Then, if $\kappa > \bar{\kappa}(f)$, the principal offers $C_F(f)$ with probability one.

Now, suppose that $\kappa \leq \bar{\kappa}$. The principal must now mix between $C_F(f)$ and $C_R(f)$. Let $\eta^*(f)$ be the equilibrium probability that the principal chooses contract $C_F(f)$. In order to be willing to mix, the principal must be indifferent between offering the two contracts which implies that $\rho(f)$ must be equal to $D(f_P||P)/(1 + D(f_P||P))$. Given $\eta^*(f)$, the optimal choice of $\rho$ solves

$$\max_{\rho \in [0,1]} -\eta^*(f)(1-\rho)\frac{\theta^2}{c}D(f_P||P) - T(\rho)$$

which $D(f_P||P)/(1+D(f_P||P)$ as the solution to the first-order condition if and only if

$$\eta^*(f)\frac{\theta^2}{\kappa c}D(f_P||P) = \frac{D(f_P||P)}{1 + D(f_P||P)} \Leftrightarrow \eta^*(f) = \frac{\kappa}{\bar{\kappa}(f)}.$$

Hence, for $\kappa \leq \bar{\kappa}(f)$, $\eta^*(f) = \kappa/\bar{\kappa}(f)$ gives the probability that contract $C_F(f)$ is offered in equilibrium. ■

### A.1.3 Proof of Proposition 2.1

Suppose that $\kappa \leq \bar{\kappa}(f)$. From Lemma 2.2 we know that the principal must be indifferent between offering the contract $C_F(f)$ and $C_R(f)$ in this region, which is only the case if $\rho(f) = D(f_P||P)/(1 + D(f_P||P)$. Hence, this is the equilibrium cognitive strategy of the agent for $\kappa$ sufficiently small.

Now, suppose that $\kappa > \bar{\kappa}(f)$. Then, the principal offers only the contract $C_F(f)$ in equilibrium and the optimal cognitive strategy of the agent is given in equation

(A.2). Re-stating this expression, we have

$$\rho(f) = \frac{\theta^2}{\kappa c} D(f_P||P)$$

for $\kappa > \bar{\kappa}(f)$.

Since $D(f_P||P)$ is a divergence measure, we know that $D(f_P||P) \geq 0$ with equality if and only if $f_P = P$. Hence, $\rho(f) = 0$ only when $f_P = P$ (the principal utilizes a frame that does not impact on the agent's beliefs). It is also clear that, as $D(f_P||P)$ increases, $\rho(f)$ is increasing, irrespective of the magnitude of $\kappa$. ■

### A.1.4 Proof of Proposition 2.2

For $\kappa \leq \bar{\kappa}(f)$, the cognitive strategy of the agent is given by $\rho(f) = D(f_P||P)/(1 + D(f_P||P))$, which makes the principal indifferent between offering $C_F(f)$ and $C_R(f)$. If $C_R(f)$ is offered, the principal makes profit $\theta^2/(2c)$ on each contract sold, and both types participate. Hence, the principal's profits are equal to $\theta^2/(2c)$ for $\kappa \leq \bar{\kappa}(f)$.

Now, suppose that $\kappa > \bar{\kappa}(f)$. Now, the principal offers only $C_R(f)$ (Lemma 2.2). As computed in the proof of Lemma 2.2, the profit on each sale of contract $C_F(f)$ is given by $\frac{\theta^2}{2c}(1 + D(f_P||P))$, which is obviously strictly increasing in $D(f_P||P)$. Since the rational-type does not purchase this contract, aggregate profits are given by this value multiplied by $1 - \rho(f)$ where $\rho(f) = \frac{\theta^2}{\kappa c} D(f_P||P)$. Thus, the only way that the frame $f$ enters the objective function of the principal is through the term $D(f_P||P)$ and, as such, this divergence measure serves as a sufficient statistic for the principal's optimization problem. ■

### A.1.5 Proof of Proposition 2.3

To prove $(a)$, fix a frame and note that for $\kappa \leq \bar{\kappa}(f)$, the profits of the principal are constant, given by $\theta^2/c$ (Proposition 2.2). This condition is satisfied for all frames, $f$, if and only if it is satisfied for $D(f_P||P) = 0$, which is equivalent to $\kappa \leq \theta^2/c \equiv \underline{\kappa}$. Hence, if $\kappa \leq \underline{\kappa}$, the principal's profits are constant across all frames and, as such, any $f_P \in \mathcal{F}$ is optimal.

Now suppose that $\kappa > \underline{\kappa}$. It is now the case that there exists a frame, $f$, such that the $\kappa > \bar{\kappa}(f)$, and the principal faces the following trade-off: make the frame more exploitative at the cost of less framed types in the market. The (unconstrained) optimization problem of the principal is to choose the divergence measure that solves

$$\max_d \left(1 - \frac{\theta^2}{\kappa c} d\right) \frac{\theta^2}{2c}(1+d).$$

This leads to the first-order condition

$$-\frac{\theta^2}{\kappa c}(1+d^*) + 1 - \frac{\theta^2}{\kappa c} d^* = 0$$

which, we re-arranged, gives

$$d^* = \frac{1}{2}\left(\frac{\kappa c}{\theta^2} - 1\right).$$

Note that this is strictly greater than zero as $\kappa > \theta^2/c$, and so is a feasible solution only if it is less that $\bar{D}$, which is the case if and only if

$$\frac{1}{2}\left(\frac{\kappa c}{\theta^2} - 1\right) \leq \bar{D} \Leftrightarrow \kappa \leq \frac{\theta^2}{c}\left(2\bar{D} + 1\right).$$

If $\kappa$ exceeds this threshold, then we have that the principal chooses $f$ so that $D(f_P||P) = \bar{D}$.

The final thing to show is that the optimal choice of frame, $f^*$, is such that $\rho(f^*) < D(f_P^*||P)/(1 + D(f_P^*||P))$, which is equivalent to showing that $\kappa > \bar{\kappa}(f^*)$. This is the case if

$$\kappa > \frac{\theta^2}{c}\left(1 + \frac{1}{2}\left(\frac{\kappa c}{\theta^2} - 1\right)\right) \Leftrightarrow \kappa > \frac{\theta^2}{c}$$

which holds in this region of the parameter space. It follows that the optimal choice of divergence measure is feasible. ■

### A.1.6 Proof of Proposition 2.4

We start by proving $(a)$: comparative statics for the equilibrium cognitive strategy of the agent. Since we are imposing the equilibrium selection criterion that $f_P = P$ if the principal is indifferent between all frames, it follows that $\rho(f^*) = 0$ for all $\kappa \leq \underline{\kappa}$.

Now, suppose that $\kappa \in (\underline{\kappa}, \bar{\kappa})$. Then, we have that the optimal frame is such that

$$D(f||P) = \frac{1}{2}\left(\frac{\kappa c}{\theta^2} - 1\right).$$

The equilibrium cognitive strategy of the agent given such a frame is equal to

$$\rho(f) = \frac{\theta^2}{\kappa c} \times \frac{1}{2}\left(\frac{\kappa c}{\theta^2} - 1\right) = \frac{1}{2}\left(1 - \frac{\theta^2}{\kappa c}\right).$$

This is obviously increasing in $\kappa$ and is decreasing in $\theta^2/c$.

Now, suppose that $\kappa \geq \bar{\kappa}$. From Proposition 2.3, the optimal frame is such that $D(f||P) = \bar{D}$ which is constant in both $\kappa$ and $\theta^2/c$. It follows that

$$\rho(f) = \frac{\theta^2}{\kappa c}\bar{D}$$

which is obviously decreasing in $\kappa$ and increasing in $\theta^2/c$.

We now move on to show comparative statics for total surplus, defined as the (unweighted) sum of the principal's profits and the ex-ante welfare of the agent. To proceed, we first show that total surplus is maximized at the point in which $f_P = P$; that is, the principal does not frame the agent. This is obviously true since the rational benchmark maximizes surplus in every frame and, since the principal offers the rational contract when $f_P = P$, the result follows. ■

### A.1.7 Proof of Proposition 2.5

First, we note that the cognitive strategy of the agent is unchanged from before. The only difference is that the principal now earns profits on the rational type of the agent. Hence, the optimization problem of the principal is to find the $d^*$ that solves

$$\max_{d\in[0,\bar{D}]} \rho(d)\frac{\theta^2}{2c} + (1-\rho(d))\frac{\theta^2}{2c}(1+d).$$

Solving this gives $d^* = \frac{\kappa c}{2\theta^2}$ which is strictly greater than the equilibrium degree of exploitation derived in the baseline version of the model. Substituting this value into the expression for the agent's equilibrium cognitive strategy gives $\rho(d^*) = 1/2$ which again, is obviously strictly greater than the value derived in the baseline version of the model.

Proving the result regarding the total surplus is obvious once one observes the following two facts: (1) for $\kappa \leq \underline{\kappa}$ the model with no screening is efficient while there exist distortions in the model with screening; and (2) for $\kappa \geq \bar{\kappa}$ the principal chooses the maximally exploitative frame in both environments. Moreover, when $\kappa \geq \bar{\kappa}$, the principal is able to earn profits on the rational type, while such types do not participate in the market in the setting with a single contract. Hence, total surplus is higher for $\kappa \geq \bar{\kappa}$ when screening is permitted. Finally, by the intermediate value theorem, the desired threshold exists. ■

## A.2 Proof of Results in Chapter 3

First, I prove two useful results for the most general formulation of the model. All results concerning feasible implementation (Proposition 3.1 and Proposition 3.3) are corollaries of these lemmas.

**Lemma A.1.** *For any goal $g = (\xi, \phi)$, the best responses of self-1 satisfy*

$$e(c, g) \in [\underline{e}_c, \bar{e}_c] \equiv \left[\mu(0)\beta\frac{V}{c}, \mu(1)\beta\frac{V}{c}\right] \tag{A.3}$$

*for all $c \in \Theta$. Thus, given Assumption 3.3, $e(c, g)$ is strictly decreasing in $c$ and*

$$C(e(c_1, g), c_1) > C(e(c_2, g), c_2) > \ldots > C(e(c_n, g), c_n) \tag{A.4}$$

*Proof.* Suppose that self-0 sets an arbitrary goal, $g = (\xi, \phi)$. Suppose that cost-state $c \in \Theta$ realizes. Define the following sets, which are a function of state $c$, the effort level chosen in state $c$, and the goal $g$:

$$\nu_c^1(e, g) = \bigcup_{c' \in \xi : \phi(c') > e} \{c'\}$$

$$\nu_c^2(e, g) = \bigcup_{c' \in \xi : \phi(c') < e} \{c'\}$$

$$\nu_c^3(e, g) = \bigcup_{c' \in \xi : C(e,c) < C(\phi(c'), c')} \{c'\}$$

$$\nu_c^4(e, g) = \bigcup_{c' \in \xi : C(e,c) > C(\phi(c'), c')} \{c'\}$$

Then, for a given choice of effort, $e$, in cost-state $c$, self-1 utility, $U_1(e|c,g)$, is given by

$$\beta V e\left(1+\eta P[\nu_c^2(e,g)|\xi]+\eta\lambda P[\nu_c^1(e,g)|\xi]\right)-c\frac{e^2}{2}\left(1+\eta P[\nu_c^3(e,g)|\xi]+\eta\lambda P[\nu_c^4(e,g)|\xi]\right)+\kappa$$

where $\kappa$ is a constant term, independent of $e$. Note that this function is continuous in $e$. Moreover, the function is differentiable except at 'kink' points where it holds that either $P[\nu_c^1(e,g)|\xi]+P[\nu_c^2(e,g)|\xi]<1$ or $P[\nu_c^3(e,g)|\xi]+P[\nu_c^4(e,g)|\xi]<1$. Since there are finitely many points at which the function is non-differentiable (as the state-space is finite), to prove the first part of the lemma, it is sufficient to show that the derivative of the utility function in $e$ is strictly negative (positive) for all $e>\mu(1)\beta V/c$ ($e<\mu(0)\beta V/c$) at all points of differentiability.

Suppose that the function is differentiable at $e$ or, equivalently, $e$ is such that $P[\nu_c^1(e,g)|\xi]+P[\nu_c^2(e,g)|\xi]=1$ and $P[\nu_c^3(e,g)|\xi]+P[\nu_c^4(e,g)|\xi]=1$. At such a point, $P[\nu_c^m(e,g)|\xi]$, $m=1,2,3,4$, is constant in $e$ and the derivative of self-1 utility is given by

$$\beta V\left(1+\eta P[\nu_c^2(e,g)|\xi]+\eta\lambda P[\nu_c^1(e,g)|\xi]\right)-c\left(1+\eta P[\nu_c^3(e,g)|\xi]+\eta\lambda P[\nu_c^4(e,g)|\xi]\right)e.$$

This is strictly negative at $e$ if and only if

$$\frac{1+\eta P[\nu_c^2(e,g)|\xi]+\eta\lambda P[\nu_c^1(e,g)|\xi]}{1+\eta P[\nu_c^3(e,g)|\xi]+\eta\lambda P[\nu_c^4(e,g)|\xi]}\frac{\beta V}{c}<e.$$

Since it holds that $P[\nu_c^2|\xi]+P[\nu_c^1|\xi]=1$ and $P[\nu_c^3|\xi]+P[\nu_c^4|\xi]=1$ at a point of differentiability, it follows that the left-hand side of this inequality is maximized (with respect to the probabilities) at $\mu(1)\beta V/c$ (where $\mu(1)=(1+\eta\lambda)/(1+\eta)$) and is minimized at $\mu(0)\beta V/c$ (where $\mu(0)=(1+\eta)/(1+\eta\lambda)$. Hence, it follows that, at

points of differentiability, the derivative of the utility function is strictly negative for all $e > \mu(1)\beta V/c$ and is strictly positive for all $e < \mu(0)\beta V/c$. Combining this with the fact that there are finitely many points of discontinuity, it follows that $U_1(e|c,g)$ is strictly decreasing for $e > \mu(1)\beta V/c$ and is strictly increasing for $e < \mu(0)\beta V/c$. Thus, $e(c,g) \in [\underline{e}_c, \bar{e}_c]$ for all $c \in \Theta$ which proves the first part of the lemma.

Now, take an arbitrary $c_j \in \Theta$. To prove the second part of the lemma, it is sufficient to show that $e(c_j,g) > e(c_{j+1},g)$ and $C(e(c_j,g),c_j) > C(e(c_{j+1},g),c_{j+1})$ for all $j \in \{1,\ldots,n-1\}$. Moreover, since $c_j < c_{j+1}$, it follows that $C(e(c_j,g),c_j) > C(e(c_{j+1},g),c_{j+1})$ implies that $e(c_j,g) > e(c_{j+1},g)$. Using the first part of the lemma, we know that $C(e(c_j,g),c_j) \geq C(\underline{e}_{c_j},c_j)$ and $C(\bar{e}_{c_{j+1}},c_{j+1}) \geq C(e(c_{j+1},g),c_{j+1})$. Hence, $C(e(c_j,g),c_j) > C(e(c_{j+1},g),c_{j+1})$ will be guaranteed if

$$C(\underline{e}_{c_j},c_j) > C(\bar{e}_{c_{j+1}},c_{j+1}) \Leftrightarrow \sqrt{\frac{c_{j+1}}{c_j}} > \left(\frac{1+\eta\lambda}{1+\eta}\right)^2$$

which is precisely Assumption 1. It follows that both $e(c,g)$ and $C(e(c,g),c)$ are strictly decreasing in $c$. ■

**Lemma A.2.** *Suppose that self-0 sets a feasible goal $g = (\xi,\phi)$. Define $L(c,\xi) \equiv \bigcup_{c' \in \xi | c' < c} c'$. Then, for $c \in \xi$,*

$$\phi(c) \in \left[\mu(P[L(c,\xi)|\xi])\beta\frac{V}{c}, \mu(P[L(c,\xi)\cup\{c\}])\beta\frac{V}{c}\right]. \tag{A.5}$$

*The best-responses of self-1 are given by*

$$e(c,g) = \begin{cases} \phi(c) & \text{if } c \in \xi \\ \mu(P[L(c,\pi,\xi)|\xi])\beta V/c & \text{if } c \notin \xi \end{cases} \tag{A.6}$$

*Proof.* Take an arbitrary goal $g = (\xi, \phi)$. By Lemma A.1, one must only consider effort levels that satisfy $e(c, g) \in [\underline{e}_c, \bar{e}_c]$ as candidates for best-responses in all states $c \in \Theta$. It follows that, in a given state $c$, it must be the case that for all $c' \in \xi$ with $c' < c$, any feasible effort level $e$ that satisfies inequality (A.3) in state $c$ must satisfy $e < \phi(c')$ and $C(e, c) < C(\phi(c'), c')$ and for all $c' \in \xi$ with $c' > c$, the all feasible $e$ in state $c$ must satisfy $e > \phi(c')$ and $C(e, c) > C(\phi(c'), c')$. Hence, it follows that $L(c, \pi, \xi)$ is the union of all events that have salient states that require strictly higher effort (at strictly higher cost) than any feasible level of effort in state $c$.

Now suppose that $c \in \xi$ realizes. Then, the goal must satisfy rational expectations in this state; that is $e(c, g) = \phi(c)$. Suppose self-1 selects effort $e$ that satisfies $\phi(c) < e \leq \mu(1)\beta V/c$. Then, self-1 utility is given by

$$\begin{aligned} &\beta V e \left(1 + \eta\lambda P[L(c,\xi)|\xi] + \eta(1 - P[L(c,\xi)|\xi])\right) \\ &- c\frac{e^2}{2}\left(1 + \eta P[L(c,\xi)|\xi] + \eta\lambda(1 - P[L(c,\xi)|\xi])\right) + \kappa \end{aligned}$$

where $\kappa$ is a constant term, independent of $e$. This is strictly decreasing in $e$ for all $e > \phi(c)$ if and only if

$$\mu(P[L(c,\xi)|\xi])\beta\frac{V}{c} \equiv \frac{1 + \eta\lambda P[L(c,\xi)|\xi] + \eta(1 - P[L(c,\xi)|\xi])}{1 + \eta P[L(c,\xi)|\xi] + \eta\lambda(1 - P[L(c,\xi)|\xi])}\frac{\beta V}{c} \leq \phi(c)$$

which is a lower bound on the goal set for state $c \in \xi$. To derive an upper bound, suppose that self-1 selects effort that satisfies $\mu(0)\beta V/c \leq e < \phi(c)$. Then, self-1 utility is given by

$$\begin{aligned} &\beta V e \left(1 + \eta\lambda P[L(c,\xi) \cup \{c\}|\xi] + \eta(1 - P[L(c,\xi) \cup \{c\}|\xi])\right) \\ &- c\frac{e^2}{2}\left(1 + \eta P[L(c,\xi) \cup \{c\}|\xi] + \eta\lambda(1 - P[L(c,\xi) \cup \{c\}|\xi])\right) + \kappa \end{aligned}$$

where $\kappa$ is again independent of $e$. It is then simple to show that, in this region, utility is strictly increasing in $e$ if and only if

$$\mu(P[L(\xi_k, \xi) \cup \{c\}|\xi])\beta\frac{V}{c} \geq \phi(c)$$

which is the relevant upper bound on the goal set for cell $c \in \xi$. Combining both results gives the inequalities in equation (A.5) and the construction of these bounds ensures $e(c, g) = \phi(c)$.

Now suppose that $c \notin \xi$ realizes. Then, self-1 utility for feasible effort $e \in [\underline{e}_c, \bar{e}_c]$ is given by

$$\begin{aligned}\beta V e - c\frac{e^2}{2} + \eta\left[\beta V e(1 - P[L(c,\xi)|\xi]) - c\frac{e^2}{2}P[L(c,\xi)|\xi]\right] \\ + \eta\lambda\left[\beta V e P[L(c,\xi)|\xi] - c\frac{e^2}{2}(1 - P[L(c,\xi)|\xi])\right] + \kappa\end{aligned}$$

where $\kappa$ is a constant, independent of $e$. The first-order condition gives the unconstrained optimal $e$ is state $c$ to be $e^* = \mu(P[L(c, \xi)|\xi])\beta V/c$. Since the objective function is strictly concave and $e^* \in [\underline{e}_c, \bar{e}_c]$, it follows that $e(c, g) = \mu(P[L(c, \xi)|\xi])\beta V/c$ for all non-salient states, $c \notin \xi$. This completes the proof. ■

### A.2.1 Proof of Proposition 3.1

Set $\pi = \{\{c_L\}, \{c_H\}\}$, for which $\xi = \{c_L, c_H\}$ trivially. The fact that the set of implementable goals in each cost-state forms an interval follows directly from the fact that both states are salient and Lemma A.2. For the characterization of these intervals, we have $L(c_L, \pi, \xi) = \emptyset$ and $L(c_H, \pi, \xi) = \{c_L\}$. By Lemma A.2, the greatest lower bound on the set of feasible goals for $c_L$ is $\underline{g}_L \equiv \mu(P[\emptyset])\beta V/c_L = \mu(0)\beta V/c_L$ and the least upper bound is $\bar{g}_L \equiv \mu(P[\{c_L\}])\beta V/c_L = \mu(q)\beta V/c_L$. Similarly, the greatest

lower bound on a feasible goal for $c_H$ is $\underline{g}_H \equiv \mu(P[\{c_L\}])\beta V/c_H$ and the least upper bound is given by $\bar{g}_H \equiv \mu(P[\Theta])\beta V/c_H = \mu(1)\beta V/c_H$.

To show that ex-ante optimal effort satisfies these constraints if and only if $q \geq 1/2$ and $\beta = \mu(1-q)$, first suppose the latter holds. Note that $q \geq 1/2$ implies that $\mu(1-q) \leq 1$, so $\beta = \mu(1-q)$ is feasible. In this case, $\bar{g}_L = \mu(q)\mu(1-q)V/c_L = V/c_L$ since it is easily verifiable that $\mu(q)\mu(1-q) = 1$. Similarly, $\underline{g}_H = \mu(q)\mu(1-q)V/c_H = V/c_H$. Hence, $(V/c_L, V/c_H)$ is feasible and ex-ante optimal effort is implementable.

To prove the remaining direction, suppose that $q < 1/2$. Then, $\bar{g}_L = \beta\mu(q)V/c_L$ and, since $\beta \leq 1$ and $\mu(q) < 1$ for $q < 1/2$, it follows that $\bar{g}_L < V/c_L$ and ex-ante optimal effort is never implementable in $c_L$. If instead $q \geq 1/2$ but $\beta \neq \mu(1-q)$, then suppose that $\beta > \mu(1-q)$. It follows that, for all feasible $g_H$,

$$g_H \geq \underline{g}_H = \mu(q)\beta\frac{V}{c_H} > \mu(q)\mu(1-q)\frac{V}{c_H} = \frac{V}{c_H}$$

so that $V/c_H \notin \left[\underline{g}_H, \bar{g}_H\right]$. A symmetric argument establishes that if $q \geq 1/2$ but $\beta < \mu(1-q)$, then $\bar{g}_L < V/c_L$ so that $V/c_L \notin \left[\underline{g}_L, \bar{g}_L\right]$. Thus, $(V/c_L, V/c_H)$ is implementable only if $q \geq 1/2$ and $\beta = \mu(1-q)$, which concludes the proof. ■

### A.2.2 Proof of Proposition 3.2

Due to Proposition 3.1, the problem of self-0 is to choose $g_C^* = (g_L^*, g_H^*)$ to solve

$$\max_{g_L, g_H} q\left[Vg_L - c_L\frac{g_L^2}{2}\right] + (1-q)\left[Vg_H - c_H\frac{g_H^2}{2}\right]$$

subject to $g_L \in \left[\underline{g}_L, \bar{g}_L\right]$ and $g_H \in \left[\underline{g}_H, \bar{g}_H\right]$. Recall that ex-ante optimal effort is given by $e_0^*(c) = V/c$ for $c = c_L, c_H$. If $\beta < \mu(0)$, then $\bar{g}_j \leq \mu(1)\beta V/c_j < V/c_j$ for $j = L, H$. By strict-concavity of the objective function, it follows that $g_C^* = (\bar{g}_L, \bar{g}_H)$ is optimal in this region.

For $\beta \in [\mu(0), 1]$, define $\bar{q}(\beta) \equiv 1 - \mu^{-1}(\beta)$ where $\mu^{-1}(\cdot)$ is the inverse function of $\mu(\cdot)$. This inverse exists since $\mu(\cdot)$ is strictly increasing in $q$. It is also simple to establish $\bar{q}(\cdot)$ is strictly decreasing in $\beta$, $\bar{q}(\mu(0)) = 1$, and $\bar{q}(1) = 1/2$. Hence, $\bar{q}(\beta) \in [1/2, 1]$ for all $\beta \in [\mu(0), 1]$.

Now, take $\beta \in [\mu(0), 1]$ and $q \leq \bar{q}(\beta)$. By the definition of $\bar{q}(\cdot)$, this is equivalent to $\beta \leq \mu(1 - q)$. Under this condition, we have that $\underline{g}_H = \mu(q)\beta V/c_H \leq V/c_H$ and $\bar{g}_H = \mu(1)\beta V/c_H \geq V/c_H$, so that ex-ante optimal effort is implementable in the high-cost state. Moreover, $\bar{g}_L = \mu(q)\beta V/c_L \leq V/c_L$ with equality if and only if $\beta = \mu(1 - q)$ or $q = \bar{q}(\beta)$. By strict concavity of the objective function, we have $g_C^* = (\bar{g}_L, V/c_H)$ in this region.

Now suppose that $q > \bar{q}(\beta)$. It is simple to verify that $V/c_L \in (\mu(0)\beta V/c_L, \mu(q)\beta V/c_L)$ and $\underline{g}_H > V/c_H$. It follows that self-0 will select $g_L^* = V/c_L$ and will wish to induce as little effort in the high-cost state as possible; that is set $g_H^* = \underline{g}_H$. ■

### A.2.3 Proof of Proposition 3.3

Suppose that self-0 chooses a single salient state to write the goal: $\xi$ is a singleton. In this case, $L(\xi, \pi, \xi) = \emptyset$ trivially. By Lemma A.2, we have that

$$g_I^\xi \equiv \phi(\Theta) \in \left[\mu(P[\emptyset])\beta\frac{V}{\xi}, \mu(P[\Theta])\beta\frac{V}{\xi}\right] = \left[\mu(0)\beta\frac{V}{\xi}, \mu(1)\beta\frac{V}{\xi}\right].$$

If $\xi = c_H$, then $L(c_L, \pi, \xi) = \emptyset$. Lemma A.2 implies that $e(c_H, g_I^{c_H}) = g_I^{c_H}$ and $e(c_L, g_I^{c_H}) = \mu(P[\emptyset])\beta V/c_L = \mu(0)\beta V/c_L$.

Instead, if $\xi = c_L$, then $L(c_H, \pi, \xi) = \Theta$. Lemma A.2 implies that $e(c_L, g_I^{\xi}) = g_I^{\xi}$ and $e(c_H, g_I^{\xi}) = \mu(P[\Theta])\beta V/c_H = \mu(1)\beta V/c_H$. In this case, $\bar{e}_L \equiv \mu(1)\beta V/c_L > \mu(q)\beta V/c_L \equiv \bar{g}_L$, such that a $c_L$-salient incomplete goal can implement strictly higher effort in the low-cost state than can be achieved with a complete goal. ■

### A.2.4 Proof of Proposition 3.4

To prove $(a)$, note that the maximal level of effort that can be implemented with an incomplete goal in state $j \in \{L, H\}$ is given by $\bar{e}_j$ (Proposition 3.3). When $\beta \leq \mu(0)$, we have that $\bar{e}_j = \mu(1)\beta V/c_j \leq V/c_j \equiv e_0^*(c_j)$ with equality only when $\beta = \mu(0)$. By strict concavity of self-0's objective function, the best self-0 can do is to induce effort $\bar{e}_j$ in all $c_j$. By Proposition 3.3, the only incomplete goal that achieves this has $\xi = c_L$ and $g_I^{c_L} = \bar{e}_j$.

Suppose instead that $\beta > \mu(0)$. It is simple to verify that $V/c_j \in (\underline{e}_j, \bar{e}_j)$ for both $j = L, H$. Moreover, by Proposition 3.3, effort provision in the state that is not made salient ($c \neq \xi$) is independent of $g_I^{\xi}$. Given self-0's objective function, it follows that if she finds it optimal to make $\xi \in \Theta$ salient, she will always select $g_I^{\xi} = V/\xi$ so that ex-ante optimal effort is implemented in the salient state. Therefore, we only need to compare the utility of two goals: $g_I^{c_L} = V/c_L$ and $g_I^{c_H} = V/c_H$.

Self-0 expected utility under $g_I^{c_L} = V/c_L$ is given by

$$U_0(g_I^{c_L}) = \frac{1}{2}q\frac{V^2}{c_L} + (1-q)\mu(1)\beta\frac{V^2}{c_H}\left(1 - \frac{\mu(1)\beta}{2}\right)$$

while ex-ante utility for $g_I^{c_H} = V/c_H$ can be computed to be

$$U_0(g_I^{c_H}) = q\mu(0)\beta\frac{V^2}{c_L}\left(1 - \frac{\mu(0)\beta}{2}\right) + \frac{1}{2}(1-q)\frac{V^2}{c_H}.$$

With some manipulation, it follows that

$$U_0(g_I^{c_L}) \geq U_0(g_I^{c_H}) \Leftrightarrow LHS(q) \equiv \frac{q}{1-q}\frac{c_H}{c_L} \geq \left(\frac{1-\mu(1)\beta}{1-\mu(0)\beta}\right)^2 \equiv RHS.$$

The right-hand side of inequality (A.2.4) is strictly greater than zero for all $\beta > \mu(0)$, and is takes maximal value $\mu(1)^2$ when $\beta = 1$. Since the left-hand side of the inequality is equal to zero when $q = 0$, strictly increasing in $q$, and equal to $c_H/c_L$ when $q = 1/2$, with Assumption 3.3 it follows that there exists a unique threshold $\bar{q}_I(\beta) \in (0, 1/2)$ such that $LHS(\bar{q}_I(\beta)) = RHS$. For $q \geq \bar{q}_I(\beta)$, $g_I^{c_L} = V/c_L$ is optimal and since $e(c_H, g_I^{c_H}) = \mu(1)\beta V/c_H > V/c_H$ (Proposition 3.3), there is over-provision of effort in the high-cost state. In contrast, for $q < \bar{q}_I(\beta)$, $g_I^{c_H} = V/c_H$ is optimal and since $e(c_L, g_I^{c_H}) = \mu(0)\beta V/c_L < V/c_L$, there is under-provision of effort in the low-cost state. ■

### A.2.5 Proof of Proposition 3.5

To show (a), suppose that $\beta \leq \mu(0)$. By Proposition 3.2, the optimal complete goal is $g_C^* = (\mu(q)\beta V/c_L, \mu(1)\beta V/c_H)$ and from Proposition 3.4, we have that $g_I^{c_L} = \mu(1)\beta V/c_L$ is the optimal incomplete goal. Hence, $e(c_H, g_C^*) = e(c_H, g_I^{c_L})$ and $e(c_L, g_C^*) < e(c_L, g_I^{c_L}) \leq V/c_L$. Strict concavity of ex-ante utility implies that the $c_L$-salient, optimal incomplete goal dominates in this region.

To establish (b), there are three cases that need to be considered: (1) $q \leq \bar{q}_I(\beta)$; (2) $q \geq \bar{q}_C(\beta)$; and (3) $q \in (\bar{q}_I(\beta), \bar{q}_C(\beta))$. In case (1), $g_I^{c_H} = V/c_H$ is the

optimal incomplete goal (Proposition 3.4) while the optimal complete goal is $g_C^* = (\mu(q)\beta V/c_L, V/c_H)$ (Proposition 3.2). By Proposition 3.3, we have that $e(c_H, g_C^*) = e(c_H, g_I^{c_H})$ and $e(c_L, g_I^{c_H}) = \mu(0)\beta V/c_L < e(c_L, g_C^*) < V/c_L$ (since $q < 1/2$). Thus, $g_I^{c_H} = V/c_H$ is strictly dominated in the region for which it is the optimal incomplete goal.

In case (2), the optimal incomplete goal is $g_I^{c_L} = V/c_L$ (Proposition 3.4) while the optimal complete goal is $g_C^* = (V/c_L, \mu(q)\beta V/c_H)$. By a similar argument to above, we have that $e(c_L, g_I^{c_L}) = e(c_L, g_C^*)$ and $e(c_H, g_I^{c_L}) > e(c_H, g_C^*) \geq V/c_H$, where the last inequality follows from the fact that $q \geq \bar{q}_C(\beta)$ is equivalent to $\beta \geq \mu(1-q)$. Thus, the complete goal gives strictly higher ex-ante utility than the incomplete goal in this region.

For case (3), Proposition 3.4 and Proposition 3.2 together imply that only the indirect utilities of $g_I^{c_L} = V/c_L$ and $g_C = (\mu(q)\beta V/c_L, V/c_H)$ need be compared. Thus, to complete the proof, it is sufficient to show that there exists a threshold, $\beta^*(q) < \mu(1-q)$, such that $\beta < \beta^*(q)$ if and only if $g_I^{c_L} = V/c_L$ strictly dominates the complete goal $g_C = (\mu(q)\beta V/c_L, V/c_H)$. We proceed by computing this threshold. The ex-ante utility of the incomplete goal $g_I^{c_L}$ is given by

$$U_0(g_I^{c_L}) = \frac{1}{2}\frac{q}{c_L}V^2 + \frac{1-q}{c_H}V^2\beta\mu(1)\left(1 - \frac{\beta\mu(1)}{2}\right)$$

and self-0 utility resulting from the complete goal $g_C = (\mu(q)\beta V/c_L, V/c_H)$ is given by

$$U_0(g_C) = \frac{q}{c_L}V^2\beta\mu(q)\left(1 - \frac{\beta\mu(q)}{2}\right) + \frac{1}{2}\frac{1-q}{c_H}V^2.$$

Algebraic manipulation gives that $U_0(g_I^{c_L}) \geq U_0(g_C)$ if and only if

$$\frac{q}{1-q}\frac{c_H}{c_L} \geq \left(\frac{1-\beta\mu(1)}{1-\beta\mu(q)}\right)^2. \tag{A.7}$$

Since $q \in (0,1)$, the left-hand side of (A.7) is strictly positive. The right-hand side of (A.7) is a strictly increasing function of $\beta$, equal to zero when $\beta = \mu(0)$, and diverges to $\infty$ as $\beta \to \mu(1-q)$. Hence, by the intermediate value theorem, there exists a unique threshold $\beta^*(q) \in (\mu(0), \mu(1-q))$ such that (A.7) holds with equality. For $\beta < \beta^*(q)$, $g_I^{c_L} = V/c_L$ is the optimal goal and for $\beta > \beta^*(q)$, $g_C = (\mu(q)\beta V/c_L, V/c_H)$ is the optimal goal. ■

**Lemma A.3.** *The expression for $\beta^*(q)$ is given by*

$$\beta^*(q) = \frac{1+\gamma(q)\sqrt{c_H/c_L}}{\mu(1)+\mu(q)\gamma(q)\sqrt{c_H/c_L}}$$

*where $\gamma(q) = \sqrt{q/(1-q)}$. This function satisfies $\lim_{q\to 0}\beta^*(q) = \lim_{q\to 1}\beta^*(q) = \mu(0)$ and is inverse U-shaped in $q$, with a peak at a point strictly less than $1/2$.*

*Proof.* First, we compute the expression for $\beta^*(q)$. Equation (A.7) implies that $U_0(g_I^{c_L}) \geq U_0(g_C^*)$ if and only if

$$-\gamma(q)\sqrt{\frac{c_H}{c_L}} \leq \frac{1-\beta\mu(1)}{1-\beta\mu(q)} \leq \gamma(q)\sqrt{\frac{c_H}{c_L}}.$$

Since $\beta < \mu(1-q)$ and $\beta > \mu(0)$, it follows that the second inequality is trivially satisfied. Rearranging the first inequality gives

$$\beta \leq \frac{1+\gamma(q)\sqrt{c_H/c_L}}{\mu(1)+\mu(q)\gamma(q)\sqrt{c_H/c_L}} \equiv \beta^*(q) \tag{A.8}$$

which is the desired expression.

Inspecting equation (A.8), we see that, as $q \to 0$, $\gamma(q) \to 0$, which implies that $\beta^*(q) \to 1/\mu(1) = \mu(0)$. Moreover, as $q \to 1$, $\gamma(q) \to \infty$ which implies

$$\beta^*(q) = \frac{1/\gamma(q) + \sqrt{c_H/c_L}}{\mu(1)/\gamma(q) + \mu(q)\sqrt{c_H/c_L}} \to 1/\mu(1) = \mu(0).$$

Now, want to show that the function is inverse U-shaped. First, note that $\beta^*(q) > \mu(0)$ if and only if $\mu(1-q) > \mu(0)$, which holds since $q \in (0,1)$. Then, since $\beta^*(\cdot)$ is continuously differentiable in $q$, it is sufficient to show there exists a unique point at which $\partial\beta^*(q)/\partial q = 0$. Computing the derivative with respect to $q$ yields

$$\frac{\partial\beta^*(q)}{\partial q} = \frac{\sqrt{\frac{c_H}{c_L}}\left[\gamma'(q)\mu(1) - \mu(q)\gamma'(q) - \mu'(q)\gamma(q)\left(1 + \gamma(q)\sqrt{\frac{c_H}{c_L}}\right)\right]}{\left(\mu(1) + \mu(q)\gamma(q)\sqrt{\frac{c_H}{c_L}}\right)^2}.$$

Define the operator $T(q) \equiv \mu(1) - \gamma(q)^2(1 + 2\gamma(q)\sqrt{c_H/c_L})$. Through tedious algebra, one can show that $\partial\beta^*(q)/\partial q = 0$ if and only if $T(q) = 0$. Since $\gamma(\cdot)$ is strictly increasing in $q$, it follows that $T(\cdot)$ is strictly decreasing in $q$. Moreover, $T(0) > 0$ and $T(1/2) < 0$ if and only if $\mu(1) - 1 - 2\sqrt{c_H/c_L} < 0$, which holds by the model's assumptions. Hence, by the intermediate value theorem, there exists a unique $q^* \in (0,1)$, $q^* < 1/2$, such that $T(q^*) = 0$, which implies that there is a unique maximizer of $\beta^*(q)$. Hence, $\beta^*(q)$ is inverse U-shaped with a peak at some point $q^* < 1/2$. ■

### A.2.6 Proof of Proposition 3.6

First, suppose that $\beta \leq \mu(0)$. By Proposition 3.5, the optimal goal is incomplete and given by $g_I^{c_L} = \beta\mu(1)V/c_L$. This implements effort $e(c_H, g_I^{c_L}) = \mu(1)\beta V/c_H$ in the high-cost state. The magnitude of deviation in the high-cost state is given by $g_I^{c_L} - e(c_H, g_I^{c_L}) = \beta\mu(1)V(1/c_L - 1/c_H)$. This expression is obviously increasing in

$\beta$ since $c_L < c_H$.

Now, suppose that $\beta \in (\mu(0), \beta^*(q))$. Again, by Proposition 3.5, the optimal goal is incomplete and given by $g_I^{c_L} = V/c_L$ which induces high-cost-state effort of $e(c_H, g_I^{c_L}) = \mu(1)\beta V/c_H$. Computing $D_H = g_I^{c_L} - e(c_H, g_I^{c_L})$ gives the stated expression in the proposition. This is obviously decreasing in $\beta$ since an increase in $\beta$ only increases effort provision in the high-cost state (and does not affect the optimal goal).

Finally, if $\beta > \beta^*(q)$, then the optimal goal is fully complete and thus, constrained to satisfy rational expectations. ■

### A.2.7 Proof of Proposition 3.7

Suppose that $\beta \leq \mu(0)$, such that the optimal goal is the incomplete goal $g_I^{c_L} = \mu(1)\beta V/c_L$. Then, we can compute $WTP(\beta)$ to be

$$WTP(\beta) = \frac{1}{2}V^2\left(\frac{q}{c_L} + \frac{1-q}{c_H}\right) - \beta\mu(1)V^2\left(1 - \frac{\beta\mu(1)}{2}\right)\left(\frac{q}{c_L} + \frac{1-q}{c_H}\right)$$

which, when appropriately factorized, gives the expression stated in the proposition. This is decreasing in $\beta$ on $(0, \mu(0)]$ as $\beta < \mu(0)$ implies that $1 > \beta\mu(1)$ such that $(1 - \beta\mu(1))^2$ is decreasing.

Now, suppose that $\beta \in (\mu(0), \beta^*(q))$. In this region, the optimal goal is incomplete and given by $g_I^{c_L} = V/c_L$. Hence, ex-ante optimal utility is achieved in cost-state $c_L$ and the willingness-to-pay is determined by how much self-0 is willing to sacrifice to

achieve ex-ante optimal effort in the high-cost state. Computing gives

$$WTP(\beta) = \frac{1}{2}V^2\frac{1-q}{c_H} - \beta\mu(1)V^2\left(1 - \frac{\beta\mu(1)}{2}\right)\frac{1-q}{c_H}$$

which, again, gives the desired expression when factorized. This is now increasing in $\beta$ as $\beta > \mu(0))$ implies that $\beta\mu(1) > 1$, such that $(1-\beta\mu(1))^2$ is increasing in $\beta$. ■

### A.2.8 Proof of Proposition 3.8

Recall that the expression for $D_H$ is given by

$$D_H = \begin{cases} \mu(1)\beta V\left(\frac{1}{c_L} - \frac{1}{c_H}\right) & \text{if } \beta \leq \mu(0) \\ \frac{V}{c_L} - \mu(1)\frac{\beta V}{c_H} & \text{if } \beta \in (\mu(0), \beta^*(q)) \\ 0 & \text{if } \beta > \beta^*(q). \end{cases}$$

To show (a), note that $\mu(1) = (1+\eta\lambda)/(1+\eta)$. It is then trivial to establish that $D_H$ is increasing in $\lambda$, increasing in $c_H$, decreasing in $c_L$, and increasing in $V$ (since $c_H > c_L$).

To prove (b), suppose that $\beta \in (\mu(0), \beta^*(q))$. Substituting in $\mu(1) = (1+\eta\lambda)/(1+\eta)$ and re-arranging $D_H$ gives

$$D_H = V\left(\frac{1}{c_L} - \frac{1+\eta\lambda}{1+\eta}\frac{\beta}{c_H}\right)$$

Obviously this expression is decreasing in $c_L$ and $\lambda$, and is increasing in $c_H$. To see that it is increasing in $V$, note that $D_H$ is linear in $V$ with a strictly positive coefficient if and only if $c_H/c_L > \beta(1+\eta\lambda)/(1+\eta)$. Since $\beta \leq 1$, Assumption 3.3 is sufficient for this to hold and $D_H$ is increasing in $V$ in this region.

Part (c) of the proposition is trivial since, for all $\beta > \beta^*(q)$, a complete goal is optimal which implies $D_H = 0$. ■

### A.2.9 Proof of Proposition 3.9

*Proof.* With $\beta < \mu(0)$, it holds that the maximal level of effort that can be implemented in any state, $\bar{e}_c = \mu(1)\beta V/c$ (from Lemma A.1), is strictly less than the ex-ante optimal level of effort, $V/c$. Given self-0's strictly concave objective function, the best she can hope to implement is $(\bar{e}_{c_1}, \ldots, \bar{e}_{c_n})$. Suppose she sets the goal $g = (\{c_1\}, \mu(1)\beta V/c_1)$. This is feasible since

$$\phi(c_1) = \mu(1)\beta\frac{V}{c_1} \in \left[\mu(0)\beta\frac{V}{c_1}, \mu(1)\beta\frac{V}{c_1}\right] = \left[\mu(P[\emptyset])\beta\frac{V}{c_1}, \mu(P[\Theta])\beta\frac{V}{c_1}\right]$$

which is the constraint imposed by Lemma A.2 on the salient state $c_1$. Hence, Lemma A.2 states that $e(c_1, g) = \mu(1)\beta V/c_1$.

For all other states $c \neq c_1$, it holds that $L(c, \pi, \xi) = \Theta$. Since none of these states are salient, Lemma A.2 implies that $e(c, g) = \mu(P[\Theta])\beta V/c = \mu(1)\beta V/c$ for all $c \neq c_1$. Since, for $c \neq c_1$, $\mu(1)\beta V/c < \mu(1)\beta V/c_1$ it follows that there is downward goal deviation in all states $c_2, c_3, \ldots, c_n$, which concludes the result. ■

### A.2.10 Proof of Proposition 3.10

To show that such a set exists, we proceed by proving that ex-ante optimal effort can be induced at $\beta = \mu(0) < 1$. Suppose such an individual uses the fully complete goal $g = (\{c_1\}, V/c_1)$. By Lemma A.2, it must be that $\phi(c_1) \in [\mu(0)\beta V/c_1, \mu(1)\beta V/c_1]$ where $\mu(1)\beta = 1$ for $\beta = \mu(0)$. Hence, $V/c_1$ is feasible. For any $c \in \Theta$, $c \neq c_1$, it holds that $L(c, \pi, \xi) = \Theta$. Hence, Lemma A.2

implies $e(c, g) = \mu(P[\Theta])\beta V/c = \mu(1)\beta V/c = V/c$. Hence, ex-ante optimal effort is implemented with this goal and $\mathcal{B}$ is non-empty.

Moreover, this is the only goal that implements the ex-ante optimal level of effort at $\beta = \mu(0)$. To see this, take an arbitrary $g'$ where there exists a $c \in \Theta$, $c \neq c_1$ such that $c \in \xi$. Then, Lemma A.2 implies that the maximal level of effort that can be induced in $c_1$ is given by $\mu(P[\Theta \setminus \{c\}|\xi])\beta V/c_1$ which is strictly less than $\mu(1)\beta V/c_1$ since $P[c] > 0$. Hence, the effort induced in $c_1$ with any other goal is strictly below the ex-ante optimum. Therefore, the unique goal that induces ex-ante optimal effort at $\beta = \mu(0)$ satisfies the two properties: $|\xi^*| \leq 2$ and $\xi \subset \{c_1, c_n\}$.

Now, take a $\beta \in (\mu(0), 1]$ such that ex-ante optimal effort is induced. Denote by $g^* = (\xi^*, \phi^*)$ this goal. Want to show that the two properties must be satisfied for $g^*$. First, it holds that $c_1 \in \xi^*$. If not, we have that $L(c_1, \xi) = \emptyset$ which, by Lemma A.2, implies that $e(c_1, g^*) = \mu(0)\beta V/c_1 < V/c_1$ for $\beta \leq 1$. Similarly, it must be the case that $c_n \in \xi^*$. Else, we have that $L(c_n, \xi) = \Theta$ which implies that $e(c_n, g^*) = \mu(1)\beta V/c_n > V/c_n$ for $\beta > \mu(0)$. Hence, $c_1, c_n \in \xi^*$.

Now, suppose that there exists another state $c_j \in \xi^*$ with $c_j \neq c_1, c_n$. By definition, we have that $\{c_1\} \cup \{c_j\} \subset L(c_n, \xi^*)$. For ex-ante optimal effort to be implementable, Lemma A.2 implies that $V/c_1 \leq \mu(P[\{c_1\}|\xi^*])\beta V/c_1$ and $V/c_n \geq \mu(P[L(c_n, \xi^*)|\xi^*])\beta V/c_n$. The conditions are equivalent to

$$\mu(1 - P[\{c_1\}|\xi^*]) \leq \beta \leq \mu(1 - P[L(c_n, \xi^*)|\xi^*]) \tag{A.9}$$

where the inequality in (A.9) can be satisfied if and only if $P[c_1] \geq P[L(c_n, \xi^*)]$ which contradicts $P[c_1] \leq P[L(c_n, \xi^*)] - P[\pi_j]$ since $P[c_j] > 0$. Hence, there cannot exist

a $c \in \xi^*$ where $c \neq c_1, c_n$ if ex-ante optimal effort is to be induced. Equivalently, $\xi^* \subset \{c_1, c_n\}$ for $\beta \in (\mu(0), 1]$. Then, since the set of salient states has at most two elements, it follows that $|\xi^*| \leq 2$ which concludes the proof. ■

### A.2.11 Proof of Proposition 3.11

For $\beta \leq \mu(0)$, the optimal goal is the incomplete goal $g_I^{c_L} = \mu(1)\beta V/c_L$ which induces effort $e(c_j, g_I^{c_L}) = \mu(1)\beta V/c_j$ for $j = L, H$. In this region it holds that $V/c_j \geq \mu(1)\beta V/c_j > \beta V/c_j = e(c_j, g_\emptyset)$. By strict concavity of self-0 utility, it follows that $g_I^{c_L}$ strictly dominates $g_\emptyset$.

Now, suppose that $\beta \in (\mu(0), 1]$ and $q < 1/2$. In this region of the parameter space, the optimal goal is either the incomplete goal $g_I^{c_L} = V/c_L$ (if $\beta < \beta^*(q)$) or the complete goal $g_C = (\mu(q)\beta V/c_L, V/c_H)$ (if $\beta > \beta^*(q)$). The indirect utility to self-0 of using a goal (as a function of $\beta$) is given by $U_0^g(\beta) \equiv \max\left\{U_0^I(\beta), U_0^C(\beta)\right\}$ where

$$U_0^I(\beta) = \left[\frac{1}{2}\frac{q}{c_L} + \frac{1-q}{c_H}\beta\mu(1)\left(1 - \frac{\beta\mu(1)}{2}\right)\right]$$

and

$$U_0^C(\beta) = \left[\frac{q}{c_L}\beta\mu(q)\left(1 - \frac{\beta\mu(q)}{2}\right) + \frac{1}{2}\frac{1-q}{c_H}\right].$$

Note that $U_0^g(\cdot)$ is a continuous function of $\beta$, which is strictly decreasing in $\beta$ for $\beta < \beta^*(q)$ and strictly increasing for $\beta > \beta^*(q)$. Let

$$U_0^\emptyset(\beta) \equiv \beta\left(1 - \frac{\beta}{2}\right)\left(\frac{q}{c_L} + \frac{1-q}{c_H}\right)$$

denote the utility to self-0 of goal-abstention. Notice that this is strictly increasing in $\beta$ on $(0, 1]$. Define the operator

$$T(\beta) \equiv 2\left(U_0^{\emptyset}(\beta) - U_0^{g}(\beta)\right)$$

where it can be easily shown that $T(\beta) > 0$ implies that goal abstention, $g_{\emptyset}$, strictly dominates both $g_I^{c_L}$ and $g_C$ and $T(\beta) < 0$ implies that some form of goal-setting dominates $g_{\emptyset}$. Want to show there exists a unique threshold, $\beta_{\emptyset} \in (\mu(0), 1)$ such that $T(\beta_{\emptyset}) = 0$, $T(\beta) > 0$ for all $\beta > \beta_{\emptyset}$, and $T(\beta) < 0$ for all $\beta < \beta_{\emptyset}$.

Suppose first that $\beta^*(q) \geq 1$ so that $g_I^{c_L}$ strictly dominates $g_C$ for all $\beta$. In this case, $U_0^g(\beta) = U_0^I(\beta)$ for all $\beta \in (\mu(0), 1]$. Note that, at $\beta = \mu(0)$, $T(\mu(0)) = -(1-\mu(0))^2(q/c_L + (1-q)/c_H) < 0$ (since $\mu(0) < 1$) and at $\beta = 1$, $T(1) = (1-\mu(1))^2(1-q)/c_H$ (since $\mu(1) > 1$). Moreover, since $U_0^{\emptyset}(\cdot)$ is strictly increasing in $\beta$ and $U_0^I(\cdot)$ is strictly decreasing in $\beta$, $T(\cdot)$ is strictly increasing in $\beta$. Then, by the intermediate value theorem there exists a unique $\beta_{\emptyset} \in (\mu(0), 1)$ such that $g_{\emptyset}$ strictly dominates $g_I^{c_L}$ when $\beta > \beta_{\emptyset}$ and $g_I^{c_L}$ strictly dominates $g_{\emptyset}$ if $\beta < \beta_{\emptyset}$.

Now, suppose that $\beta^*(q) < 1$. We proceed in steps.

**Step 1: $T(\cdot)$ is strictly concave in $\beta$ on $[\beta^*(q), 1]$**

Restrict $T$ to $[\beta^*(q), 1]$. Then, $U_0^C(\beta) \geq U_0^I$ since the complete goal strictly dominates the incomplete goal in this region. Substituting in the expressions for $U_0^{\emptyset}(\beta)$ and $U_0^g(\beta) = U_0^C(\beta)$ gives

$$T(\beta) = \beta(2-\beta)\left[\frac{q}{c_L} + \frac{1-q}{c_H}\right] - \frac{q}{c_L}\beta\mu(q)(2-\beta\mu(q)) - \frac{1-q}{c_H}.$$

This is a quadratic function in $\beta$, with the coefficient on $\beta^2$ given by

$$-\left[\frac{q}{c_L}+\frac{1-q}{c_H}\right]+\frac{q}{c_L}\mu(q)^2=-(1-\mu(q))^2\frac{q}{c_L}-\frac{1-q}{c_H}<0$$

since $\mu(q)<1$ when $q<1/2$. Hence, on $[\beta^*(q),1]$, $T(\cdot)$ is a strictly-concave quadratic function in $\beta$.

**Step 2: There exists at least one $\beta\in(\mu(0),1)$ such that $T(\beta)=0$**

At $\beta=\mu(0)$, $U_0^g(\beta)=U_0^I(\beta)$ and it has already been shown that $T(\mu(0))<0$. At $\beta=1$, the optimal goal is fully complete and $T(1)=(1-\mu(q))^2q/c_L>0$ (since $\mu(q)<1$ when $q<1/2$). Hence, at $\beta=1$, $g_\emptyset$ strictly dominates both $g_I^{c_L}$ and $g_C$. Since $T(\cdot)$ is continuous in $\beta$, by the intermediate value theorem, it follows that there exists at least one $\beta\in(\mu(0),1)$ such that $T(\beta)=0$.

Let $\beta_\emptyset=\min\{\beta|T(\beta)=0\}$. The next step shows that $\beta_\emptyset$ is the unique threshold for which $T(\beta)=0$.

**Step 3: $T(\beta)>0$ if $\beta>\beta_\emptyset$ and $T(\beta)<0$ if $\beta<\beta_\emptyset$**

Suppose first that $\beta_\emptyset\in(\mu(0),\beta^*(q)]$. For all $\beta<\beta_\emptyset$ we have that $U_0^\emptyset(\beta)<U_0^\emptyset(\beta_\emptyset)$ and $U_0^I(\beta)>U_0^I(\beta_\emptyset)$. Adding these two constraints together (and scaling by two) gives $T(\beta)<T(\beta_\emptyset)=0$. For $\beta\in(\beta_\emptyset,\beta^*(q)]$, we have that $2U_0^\emptyset(\beta)>2U_0^\emptyset(\beta_\emptyset)$ and $2U_0^I(\beta)<2U_0^I(\beta_\emptyset)$. Again, adding these constraints gives $T(\beta)>T(\beta_\emptyset)=0$. Hence, $\beta_\emptyset$ serves as the desired threshold for comparing goal-abstention to incomplete goal-setting. Now, want to show that $T>0$ on $(\beta^*(q),1]$. Take an arbitrary $\beta\in(\beta^*(q),1]$. By Step 1, $T(\cdot)$ is strictly concave in $\beta$ on this interval. Hence, for all $\alpha\in(0,1)$ we

have

$$T(\alpha\beta^*(q) + (1-\alpha)) > \alpha T(\beta^*(q)) + (1-\alpha)T(1).$$

We have shown that $T(\beta^*(q)) \geq 0$ for $\beta_\emptyset \leq \beta^*(q)$ and that $T(1) > 0$. Setting $\alpha = (1-\beta)/(1-\beta^*(q))$ we get that $T(\beta) > \alpha T(\beta^*(q)) + (1-\alpha)T(1) > 0$. Hence, $\beta_\emptyset$ is a threshold that satisfies the desired properties.

Now, suppose instead that $\beta_\emptyset \in (\beta^*(q), 1)$. Take $\beta \in (\beta_\emptyset, 1)$. Since $T$ is strictly concave, it follows that $T(\alpha\beta_\emptyset + 1 - \alpha) > \alpha T(\beta_\emptyset) + (1-\alpha)T(1)$ for all $\alpha \in (0,1)$. Letting $\alpha = (1-\beta)/(1-\beta_\emptyset)$ and noting that $T(1) > 0$ and $T(\beta_\emptyset) = 0$ gives that $T(\beta) > 0$ for all $\beta > \beta_\emptyset$.

Suppose instead that $\beta \in [\beta^*(q), \beta_\emptyset)$. Let $\alpha = (1-\beta_\emptyset)/(1-\beta)$ such that $\alpha\beta + (1-\alpha) = \beta_\emptyset$. By the strict concavity of $T$ in this region, it follows that

$$0 = T(\beta_\emptyset) = T(\alpha\beta + (1-\alpha)) > \alpha T(\beta) + (1-\alpha)T(1).$$

Then, since $T(1) > 0$, it follows that $T(\beta) < 0$ for all $\beta \in [\beta^*(q), \beta_\emptyset)$. Then, since the the operator is strictly decreasing on $[\mu(0), \beta^*(q)]$, it follows that $T(\beta) < T(\beta^*(q)) < 0$ for all $\beta \in [\mu(0), \beta^*(q)]$. Combining all these results, we get that $\beta_\emptyset$ is a threshold with the desired properties.

Now, suppose that $q > 1/2$. Suppose first that $\beta \in (\mu(0), \mu(1-q))$. Then, the complete goal $g_C = (\mu(q)\beta V/c_L, V/c_H)$ is feasible. Moreover, this strictly dominates goal abstention, which induces effort $(\beta V/c_L, \beta V/c_H)$, as $V/c_L > \mu(q)\beta V/c_L > \beta V/c_L$ for $q > 1/2$ and $g_C$ induces ex-ante optimal effort in state $c_H$. Hence, by strict concavity of self-0 utility, $g_\emptyset$ is never utilized in this region.

Suppose instead that $\beta \in [\mu(1-q), 1]$. In this region, the optimal goal is complete and is given by $g_C = (V/c_L, \mu(q)\beta V/c_H)$. Define the operator

$$T(\beta) = \beta(2-\beta)\left[\frac{q}{c_L} + \frac{1-q}{c_H}\right] - \frac{q}{c_L} - \frac{1-q}{c_H}\mu(q)\beta(2-\mu(q)\beta)$$

where it is obvious that $T(\beta) > 0$ implies that $g_\emptyset$ strictly dominates $g_C$ and $T(\beta) < 0$ implies that $g_C$ strictly dominates $g_\emptyset$. Note that $T$ is continuous in $\beta$ with $T(\mu(1-q)) < 0$ (since $g_C$ achieves ex-ante optimal utility at $\beta = \mu(1-q)$ while $g_\emptyset$ does not) and $T(1) = (1-\mu(q))^2(1-q)/c_H > 0$ (since $\mu(q) > 1$ for $q > 1/2$). Moreover, it is simple to verify that the operator is strictly increasing over this region. Hence, by the intermediate value theorem, there exists a unique threshold, $\beta_\emptyset \in (\mu(1-q), 1)$ such that $T(\beta) > 0$ for $\beta > \beta_\emptyset$ and $T(\beta) < 0$ for $\beta < \beta_\emptyset$.

Combining all results, we have that for all $q \in (0,1)$, $q \neq 1/2$, there exists a threshold $\beta_\emptyset \in (\mu(0), 1)$ such that, for all $\beta > \beta_\emptyset$, goal-abstention strictly dominates goal-setting, while for $\beta < \beta_\emptyset$, some form of goal-setting strictly dominates goal-abstention. ■.

## A.3 Proof of Results in Chapter 4

### A.3.1 Proof of Lemma 4.1

Let $\mu$ denote the receiver's belief that the sender is strategic and $\nu$ denote the belief that the sender is truthful (i.e. chooses $m = 0$ when $\omega = 0$). Suppose that a message $m = 1$ is observed. For a given cognitive strategy, $\rho$, if the receiver becomes sophisticated then she expects to learn that the sender is strategic with probability $\mu$ and is honest with probability $1 - \mu$. If she learns the sender is honest, then she chooses the action 1 to perfectly match the state of the world. Instead, if she learns the sender is strategic, she chooses her action to be the conditional probability that the state is 1 given that the sender is strategic and the probability that the sender is truthful, which is given by

$$f(1, \nu) = \frac{\pi}{\pi + (1 - \pi)(1 - \nu)}.$$

If instead the sender learns nothing from her cognitive search, then she will simply set her strategy equal to the conditional probability that the state is 1 given $(\mu, \nu)$. This is given by

$$f(\mu, \nu) = \frac{\pi}{\pi + \mu(1 - \pi)(1 - \nu)}.$$

Thus, the cognitive best-response given $m = 1$ is determined as the solution to the following optimization problem:

$$\max_{\rho} -\rho\mu f(1, \nu)(1 - f(1, \nu)) - (1 - \rho) f(\mu, \nu)(1 - f(\mu, \nu)) - \kappa \frac{\rho^2}{2}$$

subject to $\rho \in [0, 1]$. The unconstrained solution to this problem is given by

$$\rho(\mu, v) = \frac{f(\mu, \nu)(1 - f(\mu, \nu)) - \mu f(1, \nu)(1 - f(1, \nu))}{\kappa}.$$

Given the functional forms of $f(1,\nu)$ and $f(\mu,\nu)$, through tedious algebra it is possible to show that

$$f(\mu,\nu)(1-f(\mu,\nu))-\mu f(1,\nu)(1-f(1,\nu)) = \mu(1-\nu)\frac{1-\pi}{\pi}\left[f(\mu,\nu)^2 - f(1,\nu)^2\right] \equiv \bar{\kappa}(\mu,\nu).$$

It follows that the cognitive best-response function of the receiver is $\rho(\mu,\nu) = \bar{\kappa}(\mu,\nu)/\kappa$. ■

### A.3.2 Proof of Lemma 4.2

Fix a vector $(\mu,\nu)$. As shown in the proof of Lemma 4.1, if the receiver is naive then she will choose $f(\mu,\nu)$ and the sophisticated receiver will choose action $f(1,\nu)$. Given that the cognitive best-response of the agent is given by $\rho(\mu,\nu)$, it follows that the period profits of the sender conditional on $m=1$ being observed by the receiver are given by

$$\varphi(\mu,\nu) = \rho(\mu,\nu)f(1,\nu) + (1-\rho(\mu,\nu))f(\mu,\nu).$$

As $\nu$ increases, $\rho(\mu,\nu)$ decreases (Lemma 4.1) and both $f(1,\nu)$ and $f(\mu,\nu)$ increase. Since $f(\mu,\nu) > f(1,\nu)$, it follows that $\varphi$ is strictly increasing $\nu$. For $\mu$, notice that $f(\mu,\nu)$ is decreasing in $\mu$. Hence, over the region in which $\rho(\mu,\nu)$ is increasing in $\mu$, it follows that $\varphi$ is decreasing in $\mu$. When $\rho(\mu,\nu)$ is decreasing in $\mu$, it can be shown that $f(\mu,\nu)$ decreases at a rate faster than $(1-\rho(\mu,\nu))$ increases. Hence, $\varphi$ is decreasing in $\mu$. ■

### A.3.3 Proof of Lemma 4.4

Suppose $\kappa = 0$ so that the fact that the sender is strategic is observable to the receiver. Then, the unique equilibrium in $t=2$ (given in Lemma 4.3) reduces to $\varphi(\mu_2,0) = \pi$ for all $\mu_2 \in [0,1]$. If the sender tells the truth in $t=1$, then her aggregate payoff

is $0 + \pi = \pi$. Instead, if the sender lies, then she receives $x + \pi$ where $x \geq \pi > 0$. Hence, she will always have an incentive to lie and this is the unique equilibrium. ■

### A.3.4 Proof of Lemma 4.5

Suppose that $\pi \geq 1/2$. We first show that $m_1(0) = 1$ with probability one constitutes a Bayesian Nash equilibrium. To see this is the case, suppose that the sender tells the truth with probability one. It follows that his aggregate payoff is given by

$$\rho(\mu_1, 0)\pi + (1 - \rho(\mu_1, 0))f(\mu_1, 0) + \pi$$

where the final $\pi$ term is the continuation payoff that results from the fact that the $t = 2$ receiver will know with certainty that the sender is strategic. If, in contrast, the sender tells the truth ($m_1(0) = 0$), then he will receive zero payoff in $t = 1$ and a payoff of one in $t = 2$ (as the $t = 2$ receiver will believe that the sender is honest with probability one given her beliefs over the strategic sender's message strategy in $t = 1$). Since $\pi \geq 1/2$, it follows that

$$\rho(\mu_1, 0)\pi + (1 - \rho(\mu_1, 0))f(\mu_1, 0) + \pi \geq 1$$

which implies $m_1(0) = 1$ constitutes an equilibrium.

Now, want to show that this equilibrium is unique. By contradiction, suppose that there exists an equilibrium where the sender tells the truth with probability $\nu > 0$. Then, the payoff to choosing $m_1(0) = 1$ is given by

$$\rho(\mu_1, \nu)f(\mu_1, \nu) + (1 - \rho(\mu_1, \nu))f(\mu_1, \nu) + \pi$$

while the payoff to choosing $m_1(0) = 0$ is given by

$$\rho(\mu_2, 0)f(\mu_2, 0) + (1 - \rho(\mu_2, 0))f(\mu_2, \nu)$$

where $\mu_2$ is the Bayesian update of $\mu_1$ given the message m = 0 is observed in $t = 1$. It is easy to see that the payoff to choosing $m = 1$ is strictly greater than $2\pi$ which is strictly greater than one. On the contrary, the payoff to choosing $m_1(0) = 0$ is bounded above by one. Hence, the sender would wish to deviate to $m_1(0) = 1$ with probability one which is a contradiction. ■

### A.3.5 Proof of Proposition 4.1

Suppose by contradiction that $m_1(0) = 0$ with probability one constitutes an equilibrium. Since this equilibrium involves full truth-telling in $t = 1$, it follows that (a) the receiver will be naive and choose an action of zero in $t = 1$ and (b) the $t = 2$ receiver will learn nothing from observing that $m_1 = 0$ was transmitted and, as such, will hold belief $\mu_2 = \mu_1$. Hence, the payoff from telling the truth is equal to

$$\rho(\mu_1, 0)f(1, 0) + (1 - \rho(\mu_1, 0))f(\mu_1, 0).$$

Suppose instead that the sender deviates to $m_1(0) = 1$. Then, in $t = 1$, everyone is naive upon observing $m = 1$ and so will not detect the deviation. However, the $t = 2$ receiver will know that the sender is strategic with certainty. This implies the payoff to the sender from deviating is $1 + \pi$. Since 1 is a strict upper bound on the $t = 2$ payoff from telling the truth, it follows that this deviation is strictly profitable. ■

### A.3.6 Proof of Proposition 4.2

From Proposition 4.1 it holds that telling the truth with probability one can not constitute an equilibrium. Moreover, always lying can also not constitute an equilibrium in this case. To see this, note that if the sender always lies then his payoff is given by

$$\rho(\mu_1, 0)f(1, 0) + (1 - \rho(\mu_1, 0))f(\mu_1, 0) + \pi$$

while the payoff from telling the truth (if the receiver is expecting the sender to lie) is equal to one. Since $\pi < 1/2$ and $\mu_1 > \pi/(1 - \pi)$, it follows that the payoff to the sender from lying is bounded above by $2\pi < 1$ since $\pi < 1/2$. Hence, the sender has an incentive to deviate and always lying can not constitute an equilibirum.

It follows that the sender that observes $\omega_1 = 0$ must mix between sending $m = 0$ and $m = 1$. Let $\nu$ denote the probability that the sender tells the truth. Then, the payoff to choosing to lie (and send $m = 1$) is given by $\varphi(\mu_1, \nu) + \pi$. Similarly, if the sender tells the truth, then given $\nu$ the receiver will perform a Bayesian update conditional on the information content of the message, and will believe the sender is strategic with probability

$$\mu_2 = \frac{\mu_1 \nu}{\mu_1 \nu + (1 - \mu_1)}$$

in $t = 2$. It follows that the payoff from sending $m = 0$ is given by $0 + \varphi(\mu_2, 0)$ where $\mu_2$ is defined as above.

For the sender to be indifferent between sending these two messages, it must be the case that the payoffs to sending each types of message are equal. Notice that the payoff to sending $m = 1$ is strictly increasing in $\nu$ (Lemma 4.2) and the payoff to sending $m = 0$ is strictly decreasing in $\nu$ (Lemma 4.2). It is simple to check that the necessary conditions are satisfied for there to be at least one intersection, and the

monotonicity of each side ensures that this intersection is unique. Hence, the unique $\nu^*$ that solves $\varphi(\mu_1, \nu^*) + \pi = \varphi(\mu_2, 0)$ is equilibrium in this case. ■

### A.3.7 Proof of Proposition 4.3

In order to show this, it is first useful to note that, for any $\kappa > 0$, $\rho(\mu_2(\nu^*), 0)$ must be strictly less than one. If this were not the case, then the sender would receive $\pi$ in $t = 1$, irrespective of whether he lies or not in $t = 1$. Obviously this cannot be part of an equilibrium as there needs to be some investment value to telling the truth.

Recall that the equilibrium condition is determined by

$$\varphi(\mu_1, \nu^*; \kappa) + \pi = \varphi(\mu_2(\nu^*), 0; \kappa).$$

By use of the implicit function theorem, it is simple to show that $\nu^*$ is increasing in $\kappa$ (when $\rho$ is differentiable) if and only if

$$\frac{\partial \varphi}{\partial \kappa}(\mu_1, \nu^*, \kappa) > \frac{\partial \varphi}{\partial \kappa}(\mu_2(\nu^*), 0, \kappa)$$

which holds. Hence, $\nu^*$ is increasing in $\kappa$. ■

# Bibliography


Ahn, David S and Haluk Ergin. 2010. "Framing Contingencies." *Econometrica* 78 (2):655–695.

Ali, S Nageeb. 2011. "Learning Self-Control." *The Quarterly Journal of Economics* 126 (2):857–893.

Allen, Eric J, Patricia M Dechow, Devin G Pope, and George Wu. 2016. "Reference-Dependent Preferences: Evidence from Marathon Runners." *Management Science* .

Armstrong, Mark and John Vickers. 2012. "Consumer protection and contingent charges." *Journal of Economic Literature* 50 (2):477–493.

Ashraf, Nava, Dean Karlan, and Wesley Yin. 2006. "Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines." *The Quarterly Journal of Economics* :635–672.

Benabou, Roland and Guy Laroque. 1992. "Using Privileged Information to Manipulate Markets: Insiders, Gurus, and Credibility." *The Quarterly Journal of Economics* 107 (3):921–958.

Benabou, Roland and Jean Tirole. 2002. "Self-Confidence And Personal Motivation." *The Quarterly Journal of Economics* 117 (3):871–915.

Bénabou, Roland and Jean Tirole. 2004. "Willpower and Personal Rules." *Journal of Political Economy* 112 (4):848–886.

Benkert, Jean-Michel and Nick Netzer. 2016. "Informational Requirements of Nudging." .

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. "Stereotypes." *The Quarterly Journal of Economics* 131 (4):1753–1794.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2012. "Salience Theory of Choice Under Risk." *Quarterly Journal of Economics* 127 (3):1243–1285.

———. 2013. "Salience and Consumer Choice." *Journal of Political Economy* 121 (5):803–843.

Caplin, Andrew and Daniel Martin. 2012. "Framing Effects and Optimization." .

Carrillo, Juan D and Mathias Dewatripont. 2008. "Promises, Promises,?" *The Economic Journal* 118 (531):1453–1473.

Carrillo, Juan D and Thomas Mariotti. 2000. "Strategic Ignorance as a Self-Disciplining Device." *The Review of Economic Studies* 67 (3):529–544.

Chen, Ying. 2011. "Perturbed Communication Games with Honest Senders and Naive Receivers." *Journal of Economic Theory* 146 (2):401–424.

Clary, E Gil and Abraham Tesser. 1983. "Reactions to Enexpected Events: The Naive Scientist and Interpretive Activity." *Personality and Social Psychology Bulletin* 9 (4):609–620.

Crawford, Vincent P and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica: Journal of the Econometric Society* :1431–1451.

Della Vigna, Stefano and Ulrike Malmendier. 2006. "Paying Not to Go to the Gym." *The American Economic Review* 96 (3):694–719.

DellaVigna, Stefano and Ulrike Malmendier. 2004. "Contract design and self-control: Theory and evidence." *The Quarterly Journal of Economics* 119 (2):353–402.

Dupas, Pascaline and Jonathan Robinson. 2013. "Savings Constraints and Microenterprise Development: Evidence from a Field Experiment in Kenya." *American Economic Journal: Applied Economics* 5 (1):163–192.

Filiz-Ozbay, Emel. 2012. "Incorporating Unawareness into Contract Theory." *Games and Economic Behavior* 76 (1):181–194.

Gabaix, Xavier and David Laibson. 2006. "Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets." *The Quarterly Journal of Economics* 121 (2):505–540.

Gennaioli, Nicola and Andrei Shleifer. 2010. "What Comes to Mind." *The Quarterly Journal of Economics* 125 (4):1399–1433.

Hastie, Reid. 1984. "Causes and Effects of Causal Attribution." *Journal of Personality and Social Psychology* 46 (1):44.

Heath, Chip, Richard P Larrick, and George Wu. 1999. "Goals as Reference Points." *Cognitive Psychology* 38 (1):79–109.

Heidhues, Paul and Boton Koszegi. 2010. "Exploiting naivete about self-control in the credit market." *The American Economic Review* 100 (5):2279–2303.

Hsiaw, Alice. 2013. "Goal-Setting and Self-Control." *Journal of Economic Theory* 148 (2):601–626.

———. 2016. "Goal Bracketing and Self-Control." .

Jain, Sanjay. 2009. "Self-Control and Optimal Goals: A Theoretical Analysis." *Marketing Science* 28 (6):1027–1045.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. Macmillan.

Kahneman, Daniel and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47 (2):263–292.

———. 1981. "The Framing of Decisions and the Psychology of Choice." *SCIENCE* 211:30.

Kartik, Navin, Marco Ottaviani, and Francesco Squintani. 2007. "Credulity, Lies, and Costly Talk." *Journal of Economic Theory* 134 (1):93–116.

Koch, Alexander K and Julia Nafziger. 2011. "Self-Regulation through Goal Setting." *The Scandinavian Journal of Economics* 113 (1):212–227.

Kőszegi, Botond. 2014. "Behavioral contract theory." *Journal of Economic Literature* 52 (4):1075–1118.

Kőszegi, Botond and Matthew Rabin. 2006. "A Model of Reference-Dependent Preferences." *The Quarterly Journal of Economics* :1133–1165.

Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." *The Quarterly Journal of Economics* :443–477.

Locke, Edwin A and Gary P Latham. 2002. "Building a Practically Useful Theory of Goal Setting and Task Motivation: A 35-Year Odyssey." *American Psychologist* 57 (9):705.

Markle, Alex, George Wu, Rebecca J White, and Aaron M Sackett. 2015. "Goals as Reference Points in Marathon Running: A Novel Test of Reference Dependence." *Fordham University Schools of Business Research Paper* (2523510).

Morris, Stephen. 2001. "Political Correctness." *Journal of Political Economy* 109 (2):231–265.

Mullainathan, Sendhil, Joshua Schwartzstein, and Andrei Shleifer. 2008. "Coarse Thinking and Persuasion." *The Quarterly Journal of Economics* 123 (2):577–619.

Ottaviani, Marco and Francesco Squintani. 2006. "Naive Audience and Communication Bias." *International Journal of Game Theory* 35 (1):129–150.

Phelps, Edmund S and Robert A Pollak. 1968. "On Second-Best National Saving and Game-Equilibrium Growth." *The Review of Economic Studies* 35 (2):185–199.

Piccione, Michele and Ran Spiegler. 2012. "Price Competition Under Limited Comparability." *The Quarterly Journal of Economics* 127 (1):97–135.

Salant, Yuval and Ariel Rubinstein. 2008. "(A, f ): Choice with Frames." *Review of Economic Studies* :1287–1296.

Salant, Yuval and Ron Siegel. 2016. "Contracts with Framing." .

Sobel, Joel. 2013. "Giving and Receiving Advice." *Advances in Economics and Econometrics* 1:305–341.

Spiegler, Ran. 2014. "Competitive Framing." *American Economic Journal: Microeconomics* 6 (3):35–58.

Suvorov, Anton and Jeroen Van de Ven. 2008. "Goal Setting as a Self-Regulation Mechanism." *SSRN Working Paper 1286029* .

Tirole, Jean. 2009. "Cognition and Incomplete Contracts." *The American Economic Review* 99 (1):265–294.

Von Thadden, Ernst-Ludwig and Xiaojian Zhao. 2012. "Incentives for Unaware Agents." *The Review of Economic Studies* 79 (3):1151–1174.

Young, Benjamin. 2017. "Goal-Setting and Endogenous Awareness." .